View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Access LMU

Bachelorarbeit

Analyse der Studierendenbefragung der Universitätsbibliothek München mithilfe von partitionierenden Clusterverfahren

Autor:

Fabian Eifler

Betreuerin:

Prof. Dr. Bettina Grün

SS 2011

Institut für Statistik
Ludwig-Maximilians-Universität München

31. Juli 2011

Zusammenfassung

Mithilfe von partitionierenden Clusterverfahren werden Daten, die aus einer Umfrage unter Nutzern der Universitätsbibliothek München gewonnen wurden, untersucht. Es wird dabei versucht, intuitiv logische Gruppen, die durch Überlegung gebildet wurden, mittels K-Zentroid Clusteranalyse wiederzufinden. Untersucht werden dabei die Variablen Lernort, Nutzung der Services, Lernzeiten, sowie Ausstattungswünsche und Nutzung der Services gemeinsam. Eine zuvor überlegte Einteilung in Nutzerprofile lässt sich am deutlichsten bei der Clusterung der Services finden. Die Verwendung der K-Zentroid Clusteranalyse stellt sich als effizient, aber schwierig zu interpretieren heraus.

Inhaltsverzeichnis

1. Einleitung **1**
1.1. Motivation . . . 1
1.2. Aufbau der Arbeit . . . 2
1.3. Intuitiv logische Gruppen . . . 2

2. Methoden **3**
2.1. Clusteranalyse . . . 3
2.1.1. Definition . . . 3
2.1.2. Verfahren . . . 3
2.2. Ähnlichkeits- und Distanzmaß . . . 4
2.2.1. Definition . . . 4
2.2.2. Beziehung zwischen Ähnlichkeit und Distanz . . . 5
2.3. Spezielle Distanzmaße . . . 6
2.3.1. Binäre Distanzmaße . . . 6
2.3.2. Ordinale Distanzmaße . . . 8
2.3.3. Gemischte Distanzmaße . . . 9
2.4. K-Zentroid Clusteranalyse . . . 9
2.5. Visualisierung von K-Zentroid Clusteranalysen . . . 12
2.5.1. Shadow-Plot . . . 13
2.5.2. Nachbarschaftsgraph . . . 13
2.6. Tests . . . 14
2.6.1. Chi-Quadrat-Unabhängigkeitstest . . . 14
2.6.2. Kruskal-Wallis-Test . . . 15

3. Datenbeschreibung **17**
3.1. Zur Clusterbildung verwendete Variablen . . . 18
3.2. Hintergrundvariablen . . . 18
3.3. Zusammengesetzte Variablen (Zufriedenheit) . . . 19

4. Analysen **21**
4.1. Lernort 22
4.2. Services 24
4.3. Lernzeiten 30
4.4. Ausstattungswünsche und Services 37

5. Schluss **41**
5.1. Zusammenfassung der Ergebnisse 41
5.2. Ausblick 42

Literaturverzeichnis **45**

Anhang **48**

A. Grafiken und Tabellen **49**

B. R Code **55**
B.1. Computational Details 55
B.2. Funktionen 55
B.3. Beispielhafter Aufruf einer Clusteranalyse 57

C. Inhalt der CD **59**

1. Einleitung

1.1. Motivation

„Der da drüben sieht aus wie ein typischer BWLer!“ Solche Aussagen hat eigentlich jeder Studierende schon einmal gehört und es könnte anstatt Betriebswirtschaftslehre auch jedes anderes Fach sein. Solche Vorurteile sind Versuche, eine Person anhand von äußeren Kriterien einer größeren Gruppe zuzuordnen.
Generell versucht man schon seit Urzeiten, Menschen in Typen (=Gruppen) zu unterteilen. Die Hindus verwendeten z.B. Geschlecht, physische Merkmale ebenso wie Verhaltensmerkmale, um Menschen in sechs Typen zu unterteilen, denen sie die Namen von Tieren gaben (vgl. Everitt 1974).
Ähnliches gibt es auch im europäischen Raum. Die Griechen der Antike unterschieden Menschen in vier verschiedene Temperamente. Der Psychologe Kretschmer prägte davon ausgehend die vier Konstitutionstypen Leptosom, Pykniker, Sanguiniker und Athlet, wobei er äußeres Erscheinungsbild und Reaktionsweisen zusammen einer Gruppe zuordnete (vgl. Wikipedia 2010).

Die vorliegende Arbeit beschäftigt sich nicht mit der Frage, ob man Studierende anhand ihres äußeren Erscheinungsbildes klassifizieren kann, sondern ob man mit statistischen Methoden Gruppen bilden kann, die intuitiv logisch erscheinen.
Die Daten zur Beantwortung dieser Frage stammen aus einer Umfrage der Universitätsbibliothek der Ludwig-Maximilians-Universität München (im Folgenden nur noch mit UB abgekürzt) im Herbst 2010. In dieser nicht-repräsentativen Umfrage wurden von Studierenden der LMU einige allgemeine Angaben, sowie vor allem Angaben zum Lernverhalten und zur Zufriedenheit mit den verschiedenen Services der UB abgefragt. Die Dokumentation der Umfrage und ihre deskriptiven Ergebnisse univariater Analysen kann in Eifler u. a. (2011) nachgelesen werden.

1.2. Aufbau der Arbeit

Zunächst werden einige dieser intuitiv logischen Gruppen in Abschnitt 1.3 beschrieben. Anschließend werden in Kapitel 2 die verwendeten Verfahren erläutert. Außerdem werden die dafür benötigten Distanzmaße beschrieben.
In Kapitel 3 werden die Variablen beschrieben, die im anschließenden Analyseteil (Kapitel 4) verwendet werden. Abschließend werden in Kapitel 5 die Ergebnisse zusammengefasst und ein Ausblick gegeben.

1.3. Intuitiv logische Gruppen

Untersucht wird, ob sich folgende Gruppen hinsichtlich allgemeiner Variablen (Alter, Geschlecht, Anzahl Hochschulsemester, Studienfach) oder ihres Antwortverhaltens in Bezug auf die Zufriedenheit mit der UB unterscheiden:

- Studierende, die sich bezüglich ihres Lernverhaltens (Lernort und Lernzeit) ähneln.
- Studierende, die ähnlich intensiv die Services der UB nutzen.
- Studierende, die sich hinsichtlich der Nutzung der Services und der Ausstattungswünsche für ihre Lernumgebung ähneln.

2. Methoden

2.1. Clusteranalyse

2.1.1. Definition

Zum Begriff „Clusteranalyse“ findet man in beinahe jedem Buch, das dieses Thema behandelt, eine eigene Definition. Diese verschiedenen Definitionen sind sich jedoch so ähnlich, dass an dieser Stelle die Definition aus Fahrmeir u. a. (1996, S. 437) verwendet werden soll:

Definition 1 *Unter Clusteranalyse versteht man Verfahren zur Klassenbildung, d.h. zur Einteilung einer Menge von Objekten in kleinere Teilmengen, wobei sich Objekte derselben Teilmenge möglichst „ähnlich“ und Objekte aus unterschiedlichen Teilmengen möglichst „verschieden“ sein sollen. Eine solche Teilmenge wird im Folgenden als Cluster bezeichnet.*

Die Begriffe „ähnlich“ und „verschieden“ bilden dabei den kritischen Punkt dieser Definition und werden von verschiedenen Autoren auch durchaus unterschiedlich präzisiert.

2.1.2. Verfahren

Unter anderem aufgrund dieser unterschiedlichen Präzisierungen haben sich im Laufe der Zeit eine Vielzahl von Verfahren und Methoden angesammelt, über die an dieser Stelle nur ein grober Überblick gegeben werden soll (siehe Everitt 1974, S.7).

- **Partitionierende Verfahren**: Hier wird versucht, Cluster in den Daten zu finden, indem ein so genanntes „Cluster-Kriterium“ minimiert wird. Ziel ist dabei, jedem Objekt eine eindeutige Clusterzugehörigkeit zuzuweisen.

- **Hierarchische Verfahren**: Hierbei wird versucht, eine Baumstruktur innerhalb der zu untersuchenden Daten zu finden.

- **Verklumpungsverfahren (engl. Clumping techniques)**: Hierbei handelt es sich meist um partitionierende Verfahren, bei denen die Clusterzugehörigkeit nicht eindeutig bestimmt ist. Ein Objekt kann zu mehreren Clustern gehören, bzw. die Cluster können sich gegenseitig „überlappen“.
- **Dichte oder Modus-suchende Verfahren (engl. Density or mode-seeking techniques)**: Hier werden Cluster geformt, indem nach Bereichen mit hoher Dichte an Objekten gesucht wird.
- **Andere**: Alle Verfahren, die nicht klar in eine der oben genannten Gruppen fallen.

Da eine Vorstellung sämtlicher Methoden den Rahmen dieser Arbeit sprengen würde, wird hier nur auf die für die späteren Analysen notwendigen, partitionierenden Verfahren genauer eingegangen.

2.2. Ähnlichkeits- und Distanzmaß

Wie bereits in Definition 1 erwähnt, wird beim Clustering versucht, große „Ähnlichkeit“ innerhalb der Cluster (oder „Distanz“ zwischen den Clustern) zu erreichen. Zunächst sind also Maße für „Ähnlichkeit“ und „Distanz“ zu definieren.

2.2.1. Definition

Ähnlichkeitsmaß (aus Fahrmeir u. a. 1996, S. 440)

Definition 2 *Sei $X = \{X_1, \ldots, X_N\}$ eine Menge von N Objekten. Die Funktion $s : X \times X \to \mathbb{R}$ heißt Ähnlichkeitsmaß, wenn $\forall i, j = 1, \ldots, N$ gilt:*

1. $s_{ij} = s_{ji}$
2. $s_{ij} \leq s_{ii}$

Außerdem fordert man häufig auch noch, dass $s_{ij} \geq 0$ und $s_{ii} = 1$ ist. Dies wird im Falle dieser Arbeit immer der Fall sein. Die symmetrische $N \times N$-Matrix $\mathbf{S} = (s_{ij})$ heißt Ähnlichkeitsmatrix.

Distanzmaß (aus Fahrmeir u. a. 1996, S. 440f.)

Definition 3 *Sei* $X = \{X_1, \ldots, X_N\}$ *eine Menge von N Objekten. Die Funktion* $d : X \times X \to \mathbb{R}$ *heißt Distanzmaß, wenn* $\forall i, j = 1, \ldots, N$ *gilt:*

1. $d_{ii} = 0$
2. $d_{ij} \geq 0$
3. $d_{ij} = d_{ji}$

Statt d_{ij} *schreibt man auch* $d(i, j)$ *oder* $d(x_i, x_j)$*. Die symmetrische* $N \times N$*-Matrix* $\mathbf{D} = (d_{ij})$ *heißt Distanzmatrix.*
Erfüllt ein Distanzmaß d auch noch die Dreiecksungleichung,

$$d_{ij} \leq d_{il} + d_{jl} \qquad \forall i, j, l = 1, \ldots, N$$

so spricht man von einem metrischen Distanzmaß.

2.2.2. Beziehung zwischen Ähnlichkeit und Distanz

Der in dieser Arbeit verwendete Algorithmus benötigt Distanzmaße zur Bildung von Clustern. Einige Sachverhalte lassen sich aber besser durch eine Ähnlichkeit ausdrücken. Es ist z.B. viel intuitiver, die Ähnlichkeit zwischen zwei Hundearten anzugeben, als eine Distanz. Um dieses Problem zu beheben, wird also eine Transformation benötigt. Dazu fordert man, dass Objekte eine umso kleinere Distanz haben, je größer ihre Ähnlichkeit ist. Üblich sind für die Transformation von Ähnlichkeitsmaßen in Distanzmaße (vgl. Fahrmeir u. a. 1996, S. 442):

$$d_{ij} = 1 - s_{ij} \qquad \text{für} \qquad 0 \leq s_{ij} \leq 1$$

und

$$d_{ij} = \sqrt{2(1 - s_{ij})} \qquad \text{für} \qquad -1 \leq s_{ij} \leq 1$$

Zu beachten ist, dass diese Transformationen nicht notwendigerweise eine metrische Distanz liefern. Die Transformation von Distanz- in Ähnlichkeitsmaße funktioniert ähnlich und kann in Fahrmeir u. a. (1996, S. 442) nachgelesen werden.

2.3. Spezielle Distanzmaße

Die in dieser Arbeit untersuchten Variablen sind überwiegend nominal (vor allem binär) sowie ordinal skaliert. Daher sollen nun die für solche Skalen geeigneten Distanzmaße erläutert werden.

2.3.1. Binäre Distanzmaße

Zunächst ist gefordert, dass die Daten in rein binärer Form vorliegen, d.h. dass jede Variable nur die Ausprägungen null oder eins annimmt. Dabei bezeichne eins das Vorhandensein eines Merkmals, sowie null das Nichtvorhandensein eines Merkmals.
Paarvergleiche zwischen zwei Objekten anhand einer Variable lassen sich nun in 4 Fälle aufteilen (vgl. Backhaus u. a. 2008, S.395f):

1. Bei beiden Objekten ist die Eigenschaft vorhanden.
2. Bei Objekt 1 ist die Eigenschaft vorhanden, bei Objekt 2 nicht.
3. Bei Objekt 2 ist die Eigenschaft vorhanden, bei Objekt 1 nicht.
4. Bei beiden Objekten ist die Eigenschaft nicht vorhanden.

Nun zählt man für alle Variablen die einzelnen Fälle und beschreibt diese wie folgt:

- a: Anzahl der postiven Übereinstimmungen
- b,c: Anzahl der Fälle 2 bzw. 3
- d: Anzahl der negativen Übereinstimmungen

		Objekt i		
		1	0	
Objekt j	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	n

Tabelle 2.1.: 2×2-Kreuztabelle für 2 Objekte mit rein binären Variablen

Jaccard-Koeffizient

Das bekannteste binäre Ähnlichkeitsmaß ist der Jaccard-Koeffizient. Dieser wird verwendet, wenn eine Asymmetrie in der Variablenstruktur vorliegt, also das gleichzeitige Vorhandensein eines Merkmals mehr über die Ähnlichkeit zweier Objekte aussagt als das Nichtvorhandensein. Das Ähnlichkeitsmaß in diesem Fall beträgt dann (z.B. Backhaus u. a. 2008, S. 397):

$$s_{ij} = \frac{a}{a+b+c}$$

Man sieht, dass negative Übereinstimmungen in dieser Funktion nicht berücksichtigt werden. In Backhaus u. a. (2008, S. 401) wird als Beispiel für die sinnvolle Anwendung des Jaccard-Koeffizienten die Nationalität genannt. So ist die Übereinstimmung im Falle des Merkmals Deutscher eben als Ähnlichkeit zu interpretieren, eine negative Übereinstimmung sagt allerdings überhaupt nichts über die Ähnlichkeit zwischen 2 Individuen aus.
Als Distanz ausgedrückt ist der Jaccard-Koeffizient (z.B. Kaufman u. Rousseeuw 2005, S. 26):

$$d_{jac}(i,j) = \frac{b+c}{a+b+c}$$

M-Koeffizient

Ein Beispiel für ein symmetrisches Ähnlichkeitsmaß ist der M-Koeffizient. Dieser misst und gewichtet die einfachen Übereinstimmungen jeweils gleich und teilt anschließend durch die Anzahl der insgesamt gemessenen Variablen. Das Maß in diesem Fall lässt sich so schreiben (z.B. Kaufman u. Rousseeuw 2005, S. 24):

$$s_{ij} = \frac{a+d}{a+b+c+d}$$

Dieses Maß ist vor allem dann sinnvoll, wenn eine positive Übereinstimmung dasselbe Gewicht wie eine negative Übereinstimmung hat. Nimmt man als Beispiel die Variable „männlich“ so ist das Vorhandensein dieses Merkmals aber auch das Nichtvorhandensein dieses Merkmals als Ähnlichkeit zu interpretieren (vgl. Backhaus u. a. 2008, S. 401).

Nominale Merkmale

Nominale Merkmale mit mehr als zwei Merkmalsausprägungen lassen sich leicht mithilfe einer Dummykodierung in mehrere binäre Merkmale umwandeln. Zu beachten ist dabei vor allem, dass negative Übereinstimmungen dadurch sehr häufig werden. Es empfiehlt sich also, ein asymmetrisches Ähnlichkeitsmaß wie den Jaccard-Koeffizienten zu verwenden. Außerdem ist zu beachten, dass bei gleichzeitiger Analyse von binären und nominalen Variablen, nominale Variable mit vielen möglichen Merkmalsausprägungen wesentlich stärker ins Gewicht fallen. Diesem Effekt kann durch eine geeignete Gewichtung entgegengewirkt werden (vgl. Backhaus u. a. 2008, S. 401f.).

2.3.2. Ordinale Distanzmaße

Zum Umgang mit ordinalen Variablen wird in Kaufman u. Rousseeuw (2005, S. 30) bemerkt, dass man sie wie nominale Variable behandeln kann, jedoch ist dies natürlich mit einem deutlichen Informationsverlust verbunden. Anders als bei vielen anderen Autoren wird vorgeschlagen, die Variablenausprägungen nicht als intervallskaliert anzunehmen, um sie dann mit einer gebräuchlichen Metrik (z.B. der euklidischen) zu behandeln. Es wird viel mehr vorgeschlagen, zuerst die zu betrachtenden Variablen anhand ihrer Ränge wie folgt zu transformieren:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}, \tag{2.1}$$

wobei r_{if} den Rang des i-ten Objekts und M_f die höchste Kategorie in der f-ten Variable bezeichnen. Durch diese Transformation wird erreicht, dass alle z_{if} zwischen null und eins liegen. Wurden die ordinalen Variablen auf diese Weise transformiert, ist es nun zulässig, die endgültige Unähnlichkeit zwischen zwei Objekten durch die Manhattan-Distanz (auch City-Block-Metrik genannt, siehe Fahrmeir u. a. 1996, S. 449) geteilt durch die Anzahl der Variablen, die keine fehlenden Werte aufweisen, zu berechnen. Diese Distanz zwischen den Objekten i und j wird im Folgenden mit $d_{ord}(i, j)$ bezeichnet.
Auf diese Art und Weise können auch problemlos ordinale Variablen mit unterschiedlicher Anzahl an Kategorien zugleich verwendet werden.

2.3.3. Gemischte Distanzmaße

Geht man noch einen Schritt weiter und versucht, gemischte Variablen zusammen zu untersuchen, so wird klar, dass keins der oben genannten Distanz- bzw. Ähnlichkeitsmaße allein passend ist. Es ist aber zulässig, die Distanz jeweils nur für Variablen mit gleichem Skalenniveau zu berechnen und anschließend eine gewichtete Summe dieser Distanzen als Distanz zwischen den jeweiligen Objekten zu verwenden (vgl. Kaufman u. Rousseeuw 2005, S. 37).
Die Distanz zwischen 2 Objekten mit einem ordinalen und einem asymmetrisch binären Teil berechnet sich demnach, in Anlehnung an das Distanzmaß von Gower (z.B. Wedel u. Kamakura 1998, S. 47), wie folgt:

$$d_{mixed}(i,j) = \frac{w_{ord} \cdot d_{ord}(i,j) + w_{jac} \cdot d_{jac}(i,j)}{w_{ord} + w_{jac}} \tag{2.2}$$

Die Gewichte w_{ord} und w_{jac} sind positive Gewichte, die je nach inhaltlicher Fragestellung sinnvoll zu wählen sind. Eine Implementierung dieses Distanzmaßes in R ist die Funktion `iDistMixed`, die im Anhang noch genauer beschrieben ist.

2.4. K-Zentroid Clusteranalyse

Partitionierende Verfahren teilen, wie schon in Abschnitt 2.1.2 beschrieben, einen Datensatz in disjunkte Teilmengen auf. Diese Teilmengen sollen in dieser Arbeit aus Studierenden bestehen, die an der Umfrage der UB teilgenommen haben. In Abschnitt 2.3.3 wurde ein Maß eingeführt, mit dessen Hilfe die Distanz zwischen zwei Objekten anhand verschiedener Variablen bestimmt werden kann. Die Idee, die nun hinter dem analytischen Verfahren steckt, ist, jede Gruppe von Studierenden durch einen durchschnittlichen Studenten darzustellen. Dieser Student ist im statistischen Sinne ein Zentroid einer solchen Teilmenge. Die Problematik dabei ist, eine Menge an Zentroiden zu finden, die den Datensatz optimal aufteilen. Jede Möglichkeit auszuprobieren, um die „perfekte Aufteilung“ zu erlangen, scheint nur für sehr kleine Datensätze realistisch. Für eine Analyse von größeren Datenmengen kommt deswegen nur ein iteratives Verfahren in Frage. Dabei wird wie folgt vorgegangen (Leisch 2006, S. 528):

1. Starte mit einer zufälligen Menge an Zentroiden C_K (z.B. durch Ziehen von K unterschiedlichen Objekten aus X_N).

2. Ordne jeden Punkt $\mathbf{x}_n \in X_N$ dem Cluster, der durch den nächsten Zentroid bestimmt ist, zu.

3. Verändere die Zentroide C_K so, dass sie jeweils den Zentroid aller ihnen zugeordneten Punkte $c(\mathbf{x}_n)$ darstellen. Dies wird folgendermaßen erreicht:

$$\mathbf{c}_k := \underset{\mathbf{c}\in C}{\operatorname{argmin}} \sum_{n:c(\mathbf{x}_n)=\mathbf{c}_k} d(\mathbf{x}_n, \mathbf{c}), \qquad k = 1, \ldots, K \tag{2.3}$$

4. Wiederhole Schritt 2 und 3 bis zur Konvergenz.

Definition 4 ***Konvergenz:*** *Sowohl Schritt 2 als auch Schritt 3 verbessern die Zielfunktion*

$$D(X_N, C_K) = \frac{1}{N} \sum_{n=1}^{N} d(\mathbf{x}_n, c(\mathbf{x}_n)) \to \min_{C_K} \tag{2.4}$$

nicht mehr.

Dabei ist $X_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ die Menge der untersuchten Objekte, K eine festgelegte Clusteranzahl und $d(\mathbf{x}_n, c(\mathbf{x}_n))$ eine Distanzfunktion. Es ist nicht garantiert, dass dieser Algorithmus gegen ein globales Optimum konvergiert und es wird deshalb vorgeschlagen, mehrere Durchgänge mit unterschiedlichen Startpartitionen durchzuführen (vgl. Leisch 2006, S. 529). Da dieser Algorithmus eine feste Anzahl Gruppen benötigt, diese in der Praxis aber oft nicht bekannt sind, wird in Wedel u. Kamakura (1998, S. 55) vorgeschlagen, mehrere Versuche mit unterschiedlicher Zentroidanzahl durchzuführen. Anschließend wird in einem Diagramm die Anzahl der Cluster und der Wert der Zielfunktion (2.4) eingezeichnet. Anhand des „deutlichsten“ Knicks in diesem Diagramm wird dann die bestmögliche Anzahl der Cluster bestimmt. Diese Art der Darstellung wird Scree-Plot genannt.
Alternativ wird vorgeschlagen, die Anzahl der Gruppen durch sachlogische Überlegungen vorzugeben (vgl. Backhaus u. a. 2008, S. 430).
Implementiert ist dieser Algorithmus in R durch die Funktion `kcca` aus dem Paket **flexclust** (Leisch 2006).

Zentroidberechnung

Wie in Abschnitt 2.4 unter Punkt 3 des Algorithmus beschrieben müssen die Zentroide immer wieder neu berechnet werden. Die Berechnung der Zentroide für die jeweiligen Skalenniveaus wird in diesem Abschnitt beschrieben.

Definition 5 ***Kanonischer Zentroid:*** *Als kanonischen Zentroid eines Clusters bezeichnet man den Punkt, der die Summe der Distanzen, aller ihm zugeordneten Punkte, zu sich minimiert. Dies entspricht der Formel 2.3.*

Ordinale Daten

Für ordinale Variablen wird nach der Transformation 2.1 eine skalierte Form der Manhattan-Distanz verwendet. Der kanonische Zentroid eines Clusters berechnet sich demnach durch den Median aller, dem Cluster zugehörigen, Beobachtungen (vgl. Leisch 2006, S. 529).

Binäre Daten

Unter Verwendung der Jaccard-Distanz für binäre Daten kann die Berechnung kanonischer Zentroide unter Umständen sehr rechenaufwändig sein. In Leisch (2006, S. 532) wird allerdings eine Lösung dieses Problems beschrieben.
Es wird ebenfalls vorgeschlagen, erwartungsbasierte Zentroide für binäre Daten zu verwenden, also das variablenweise arithmetische Mittel aller zu einem Cluster gehörenden Punkte als Zentroid zu verwenden. Dies hat neben der schnelleren Berechnung auch noch den Vorteil, dass die Zentroide als Wahrscheinlichkeiten, eine 1 zu beobachten, interpretiert werden können (vgl. Leisch 2006, S. 533). Da es sich hierbei nicht um kanonische Zentroide handelt, ist eine Konvergenz des Verfahrens nicht mehr gewährleistet. Die Empirie zeigt allerdings, dass dies in der Praxis keine Probleme verursacht (Leisch 2006, S. 533).

Gemischte Daten

Da für Daten mit Variablen verschiedener Skalenniveaus, Distanzen zunächst ebenfalls nur für gleichskalierte Variablen berechnet werden, wird bei der Zentroidberechnung ebenso verfahren. Für eine Mischung aus ordinalen und binären Daten berechnet sich der Zentroid also durch den Median der transformierten, ordinalen Variablen und dem arithmetischen Mittel der binären Variablen.
Implementiert wird diese Zentroidberechnung in R durch die Funktion `iCentMixed`, die im Anhang genauer beschrieben ist.

2.5. Visualisierung von K-Zentroid Clusteranalysen

Für die weiteren Visualisierungen der Ergebnisse einer K-Zentroid Clusteranalyse wird der so genannte Shadow-Wert benötigt.

Definition 6 *Der Shadow-Wert einer Beobachtung* $\mathbf{x}$*,* $s(\mathbf{x})$ *ist definiert als (Leisch 2010, S. 458):*

$$s(\mathbf{x}) = \frac{2d(\mathbf{x}, c(\mathbf{x}))}{d(\mathbf{x}, c(\mathbf{x})) + d(\mathbf{x}, \tilde{c}(\mathbf{x}))} \tag{2.5}$$

Dabei ist $c(\mathbf{x})$ *der zu* $\mathbf{x}$ *gehörende Zentroid und* $\tilde{c}(\mathbf{x})$ *der zweitnächste Zentroid von* $\mathbf{x}$*, im Sinne von:*

$$\tilde{c}(\mathbf{x}) = \underset{\mathbf{c} \in C_K \setminus \{c(\mathbf{x})\}}{\operatorname{argmin}} \; d(\mathbf{x}, \mathbf{c}) \tag{2.6}$$

Mit d sei die Distanzfunktion, die auch bei der Clusteranalyse verwendet wurde, bezeichnet.

Liegt der Shadow-Wert nahe bei null, so ist die Beobachtung sehr nah an ihrem Zentroid, liegt der Shadow-Wert dagegen bei fast eins, so ist die Beobachtung beinahe gleich weit von zwei Zentroiden entfernt. Der Mittelwert aller Shadow-Werte eines Clusters ist also ein Maß für die Separiertheit dieses Clusters (vgl. Leisch 2010, S. 458).
Der Shadow-Wert einer Beobachtung ist dabei eng verwandt mit dem Silhouette-Wert einer Beobachtung. Der Silhouette-Wert einer Beobachtung wird in Rousseeuw (1987) motiviert und exakt beschrieben. In Leisch (2010, S. 460) wird für den Silhouette-Wert folgende Formel angegeben:

$$sil(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max(a(\mathbf{x}), b(\mathbf{x}))}$$

Dabei ist $a(\mathbf{x})$ die durchschnittliche Distanz von $\mathbf{x}$ zu allen Werten in ihrem Cluster und $b(\mathbf{x})$ die durchschnittliche Distanz von $\mathbf{x}$ zu allen Beobachtungen in dem Cluster, der $\mathbf{x}$ am zweitnächsten liegt.
Der Hauptunterschied zwischen Silhouette-Werten und Shadow-Werten liegt dabei laut Leisch (2010, S. 460) darin, dass die durchschnittliche Distanz zu Punkten in einem Cluster, durch die Distanz zum Zentroid des Clusters ersetzt wird. Dies führt zu einer wesentlich effizienteren Berechnung.

2.5.1. Shadow-Plot

Eine Möglichkeit der grafischen Darstellung dieser Shadow-Werte bietet der Shadow-Plot. Dazu werden alle Punkte eines Clusters nach der Größe ihrer Shadow-Werte absteigend geordnet und anschließend in ein Panel gezeichnet, das in der Breite proportional zur jeweiligen Clustergröße ist. Diese Kurve wird Shadow-Kurve genannt und sollte im Idealfall so niedrig wie möglich sein. Um die verschiedenen Shadow-Kurven besser vergleichen zu können, wird außerdem noch ein Rechteck hinterlegt, welches die Fläche unter der Shadow-Kurve approximiert (vgl. Leisch 2010, S. 460).
Shadow-Plots haben den großen Vorteil, dimensionsunabhängig zu sein. Das bedeutet, dass keine Projektion notwendig ist, um einen Shadow-Plot zu erstellen, egal wieviele Variablen zur Clusterung verwendet wurden.

2.5.2. Nachbarschaftsgraph

Der Nachbarschaftsgraph bietet, mithilfe einer entsprechenden dimensionsreduzierenden Projektion, eine zweidimensionale Darstellung der Zentroide einer Clusteranalyse. Entwickelt wurde der Nachbarschaftsgraph aus den TRN-Graphen (topology-representing networks; Martinetz u. Schulten 1994) und den Silhouette-Werten. Der Nachbarschaftsgraph verwendet die Zentroide der Cluster als Knoten und verbindet diese durch gewichtete Kanten. Diese Gewichtung wird wie folgt berechnet (vgl. Leisch 2006, S. 536):
Sei $A_k \subset X_N$ die Menge aller Punkte in Cluster k,

$$A_{ij} = \{\mathbf{x}_n | \mathbf{c}_i = c(\mathbf{x}_n), \mathbf{c}_j = \tilde{c}(\mathbf{x}_n)\}$$

sei die Menge aller Punkte die $\mathbf{c}_i$ als Zentroid und $\mathbf{c}_j$ als zweitnächsten Zentroid haben.

$$e_{ij} = \begin{cases} |A_i|^{-1} \sum\limits_{\mathbf{x} \in A_{ij}} s(\mathbf{x}) & \text{für } A_{ij} \neq \emptyset \\ 0 & \text{für } A_{ij} = \emptyset \end{cases} \tag{2.7}$$

ist das Gewicht der Kante von Cluster i nach Cluster j. Dieses Gewicht ist also ein Indikator für den Zusammenhang zweier Cluster. Zur Vereinfachung der Darstellung wird statt einem gerichteten Graphen mit den oben definierten Kantengewichten ein ungerichteter Graph gezeichnet, dessen Kantengewichte die durschnittlichen Werte von e_{ij} und e_{ji} sind. Das Gewicht einer Kante ist dabei durch ihre Linienstärke dargestellt (vgl. Leisch 2006, S. 563f.).

2.6. Tests

Die folgenden Tests finden in der Statistik häufig Anwendung. Sie tauchen oftmals in der Literatur auf und die hier dargestellte Form stammt aus Duller (2008). Beide hier vorgestellten Tests führen bei hohen Fallzahlen bereits bei geringen Unterschieden zur Ablehnung der Nullhypothese.

2.6.1. Chi-Quadrat-Unabhängigkeitstest

Der χ^2-Unabhängigkeitstest testet, ob von zwei Variablen, die nominales Skalenniveau besitzen, behauptet werden kann, dass sie stochastisch unabhängig sind. Es werden dabei folgende zwei Hypothesen gegeneinander getestet:

- H_0: Die Variablen X und Y sind stochastisch unabhängig.
- H_1: Die Variablen X und Y sind nicht stochastisch unabhängig.

Bezeichne h_{ij}^o die beobachtete absolute Häufigkeit der Kombination $X = i$ und $Y = j$ mit $i = 1, \dots m$ und $j = 1, \dots, r$ Merkmalsausprägungen. Weiterhin bezeichne h_{ij}^e die unter H_0 erwartete absolute Häufigkeit dieser Kombination. Dann ergibt sich als Teststatistik:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{(h_{ij}^o - h_{ij}^e)^2}{h_{ij}^e}$$

Diese Teststatistik ist unter H_0 approximativ χ^2-verteilt. Um die Teststatistik mit Werten der χ^2-Verteilung vergleichen zu können, werden in Duller (2008) folgende Mindestanforderungen als Faustregel vorgeschlagen: Die erwartete Häufigkeit in jeder Kategorie muss mindestens 1 betragen und bei höchstens 20% der Kategorien dürfen die erwarteten Häufigkeiten unter 5 liegen. Die Nullhypothese wird mit Irrtumswahrscheinlichkeit α verworfen falls:

$$\chi^2 > \chi^2_{(m-1)(r-1);1-\alpha}$$

In R wird dieser Test durch die Funktion `chisq.test` berechnet (R Development Core Team 2010).

2.6.2. Kruskal-Wallis-Test

Um die Verteilungen einer mindestens ordinal skalierten Variablen in verschiedenen Gruppen miteinander zu vergleichen, wird der Kruskal-Wallis-Test verwendet. Mindestens ordinales Skalenniveau wird deshalb gefordert, weil die Teststatistik für den Kruskal-Wallis-Test nicht die tatsächlichen Werte, sondern die Ränge der einzelnen Beobachtungen verwendet.
Als Annahmen werden getroffen:

- Es handelt sich um unabhängige Beobachtungen.
- Die Beobachtungen folgen alle derselben unbekannten Verteilung.

Folgende Hypothesen werden geprüft:

- H_0: Die Verteilungen der Beobachtungen der einzelnen Gruppen unterscheiden sich nicht hinsichtlich ihrer Lageparameter.
- H_1: Die Verteilungen der Beobachtungen von mindestens zwei Gruppen unterscheiden sich hinsichtlich ihrer Lageparameter.

Als Teststatistik ergibt sich:

$$H = \left[\frac{12}{N(N+1)} \sum_{i=1}^{c} \frac{r_i^2}{n_i} \right] - 3(N+1)$$

Dabei ist N die Anzahl der Beobachtungen in allen Gruppen, n_i gleich der Größe der i-ten Gruppe und c gleich der Anzahl an Gruppen. Nach Zusammenfassen aller Beobachtungen zu einer gemeinsamen Gruppe, werden die Ränge der einzelnen Beobachtungen bestimmt. Mit r_i wird dann die Summe über diese Ränge in der i-ten Gruppe bezeichnet. Liegen Bindungen innerhalb einer Gruppe vor, so wird der Test nicht beeinflusst. Liegen Bindungen zwischen den Gruppen vor, so muss die Teststatistik noch mit dem Korrekturfaktor

$$C = 1 - \frac{\sum_{b=1}^{B} (l_b^3 - l_b)}{N^3 - N}$$

versehen werden. B bezeichnet die Anzahl der verschiedenartigen Bindungen zwischen den Gruppen und l_b bezeichnet, wieviele Beobachtungen in jeder dieser verschiedenartigen Bindungen enthalten sind. Als korrigierter Wert für die Teststatistik wird dann $H^* = \frac{H}{C}$ verwendet. Diese Teststatistik ist unter H_0 approximativ χ^2-verteilt. Um

die Teststatistik mit Werten der χ^2-Verteilung vergleichen zu können, werden in Duller (2008) folgende Mindestanforderungen als Faustregel vorgeschlagen: Die Gruppenumfänge n_i betragen alle mindestens fünf für c größer als drei. Die Gruppenumfänge betragen alle mindestens acht für c gleich drei. Die Nullhypothese wird mit Irrtumswahrscheinlichkeit α verworfen falls:

$$H^* > \chi^2_{c-1;1-\alpha}$$

In R wird dieser Test durch die Funktion `kruskal.test` implementiert (R Development Core Team 2010).

3. Datenbeschreibung

Die in dieser Arbeit analysierten Daten stammen, wie bereits in Abschnitt 1.1 erwähnt aus einer Onlineumfrage der UB im Wintersemester 2010/2011. Die Umfrage bestand aus drei Teilgebieten:

- Lernverhalten
- Zufriedenheit mit der UB
- allgemeine Angaben

Dabei wurden den Umfrageteilnehmern in jedem der Teilgebiete unterschiedliche Fragen gestellt, die teilweise abhängig von vorherigen Antworten und teilweise verpflichtend waren. In diesem Kapitel werden alle zur Clusterbildung und Clusterbeschreibung verwendeten Variablen kurz beschrieben. Da dies nicht alle Variablen, die durch die Umfrage erzeugt wurden, beinhaltet, sei für eine vollständige Darstellung auf Eifler u. a. (2011) verwiesen. In Eifler u. a. (2011) wurden 1602 Fragebögen als vollständig erachtet und untersucht. Im Folgenden wird jedoch die Anzahl der untersuchten Teilnehmer noch einmal reduziert. Es werden Teilnehmer ausgeschlossen, die ein Alter von mehr als 100 oder weniger als 10 Jahren angegeben haben. Ebenso werden Teilnehmer ausgeschlossen, die behaupten, mehr als 100 Hochschulsemester absolviert zu haben, da dies offensichtlich unrealistische Antworten sind und somit die Ergebnisse verfälschen würden.
Daraus resultiert eine Zahl von 1574 Umfrageteilnehmern. Im Anhang befinden sich Tabellen, in denen die Abkürzungen der verwendeten Variablen aufgelistet sind.

3.1. Zur Clusterbildung verwendete Variablen

Den Umfrageteilnehmern wurde die Frage: „Wo lernen Sie?“ gestellt und damit verlangt, zu jedem der insgesamt neun vorgegeben Lernorte anzugeben, ob sie dort „fast immer“, „häufig“, „selten“ oder „nie“ lernen. Aus dieser Frage enstehen damit neun ordinalskalierte Variablen mit den angegeben Ausprägungen.
Eine weitere Frage beschäftigte sich mit den von der UB angebotenen Services. Dabei wurden insgesamt 20 von der UB angebotene Services vorgeschlagen und Umfrageteilnehmer konnten angeben, welche dieser Services sie bereits genutzt haben. Aus dieser Frage enstehen so 20 binärskalierte Variablen, wobei eine „1“ die Nutzung eines Service kennzeichnet.
Unter der Überschrift „Zu welchen Zeiten lernen Sie?“ wurde unterteilt nach:
- Semester und Semesterferien
- wochentags, samstags, sonntags
- Tageszeit („morgens“, „mittags“, „abends“, „nachts“)
abgefragt, wann die Teilnehmer lernen. Außerdem gab es zusätzlich noch die Antwortkategorie „nie“. Mehrfachnennungen waren möglich. Daraus resultieren 30 binäre Variablen. Eine „1“ steht hier für die Angabe, zu dieser speziellen Zeit zu lernen.
Weiterhin wurde die Frage: „Was für eine Ausstattung benötigen Sie zum Lernen?“ gestellt. Die Umfrageteilnehmer konnten hier zu jedem der 11 vorgeschlagenen Objekte angeben, ob dieses für sie „wichtig“, „wünschenswert“ oder „unwichtig“ ist. Es gab die Möglichkeit, keine Angaben zu machen. Daraus resultieren 11 ordinalskalierte Variablen mit den angegebenen Ausprägungen.

3.2. Hintergrundvariablen

Der letzte Teil des Onlinefragebogens beschäftigte sich mit allgemeinen Angaben der Teilnehmer. Diese waren unkonditioniert und überwiegend verpflichtend. Damit eignen sich die Variablen dieses Teils besonders zur Beschreibung und zum Vergleich einzelner Cluster. Die erste jeweils untersuchte Variable ist das Alter der Umfrageteilnehmer. Ein Überblick der statistischen Kennzahlen ist in Tabelle 3.1, deren Visualisierung in Grafik 3.1 gegeben.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	21.00	23.00	24.17	25.00	75.00

Tabelle 3.1.: 5-Punkte-Zusammenfassung und Mittelwert des Alters

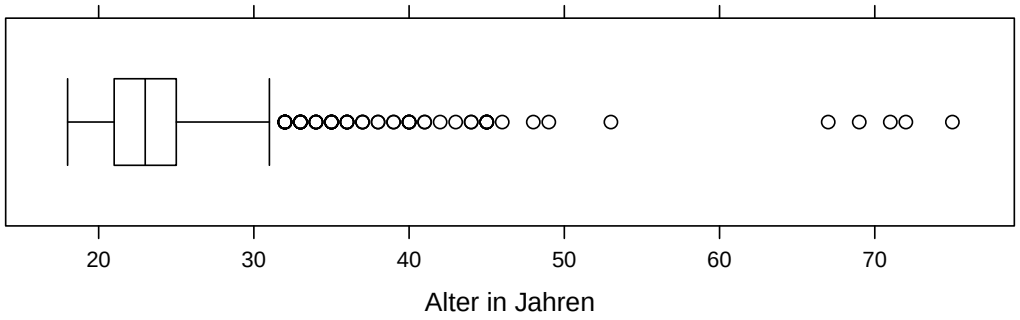

Abbildung 3.1.: Boxplot des Alters für den gesamten Datensatz

Als nächstes wurde jeweils das Geschlecht der Umfrageteilnehmer abgefragt. Der Datensatz enthält, nach den oben genannten Bereinigungen, 1061 Frauen und 493 Männer, sowie 20 Personen, die keine Angabe zu ihrem Geschlecht machten. Das Verhältnis von Frauen zu Männern beträgt demnach zwei zu eins. Grafische Darstellungen und ein Vergleich mit der Grundgesamtheit aller Studierenden der LMU kann in Eifler u. a. (2011) nachgelesen werden.
Weiterhin wurde nach dem angestrebten Abschluss und den studierten Fächern gefragt. Die Studienfächer wurden ursprünglich, mit 41 möglichen Antworten für das erste bis dritte Studienfach, sehr detailliert abgefragt. Da eine so differenzierte Abfrage des Studienfachs für die folgenden Analysen nicht notwendig ist, wird lediglich das erste Studienfach betrachtet und dieses zu den 4 Hauptfachrichtungen “Geisteswissenschaften“ (GeiWi), „Mathematik und Naturwissenschaften“ (MaNa), “Medizin und Gesundheitswissenschaften“ (Med) und „Rechts-, Wirtschafts- und Sozialwissenschaften“ (ReWiSo) zusammengefasst. Schließlich wurde noch gefragt, im wievielten Hochschulsemester sich die Teilnehmer befanden.

3.3. Zusammengesetzte Variablen (Zufriedenheit)

Aus den Fragen zum Support und zum Informationsfluss (siehe Eifler u. a. 2011, S. 30 u. 37) werden drei neue Variablen gebildet. Diese beinhalten jeweils die Anzahl an positiven, negativen und neutralen Antworten auf die insgesamt 11 Fragen zu diesen Teilgebieten. Dabei werden bei den fünf Fragen zum Support die Antwortkategorien „trifft immer zu“ und „trifft meistens zu“ als positiv, die Antwortkategorien „trifft selten zu“ und „trifft nie zu“ als negativ verstanden.

Bei den sechs Fragen zum Informationsfluss werden die Antwortkategorien „sehr gut“ und „gut“ als positiv, die Antwortkategorien „schlecht“ und „sehr schlecht“ als negativ aufgefasst. Bei beiden Fragetypen wird die Antwortkategorie „keine Antwort“ als neutral interpretiert. Diese drei für jeden Umfrageteilnehmer neu entstandenen Variablen werden im Folgenden als Beschreibung für die Zufriedenheit mit der UB insgesamt aufgefasst.

4. Analysen

Zuerst wird versucht, eine Clusteranalyse aller Teilnehmer und aller erhobenen Variablen durchzuführen. Dies scheitert an der Struktur des vorliegenden Datensatzes. Durch viele konditionierte Fragen beinhaltet jede Beobachtung einige fehlenden Werte. Der Algorithmus, der zur Clusterung verwendet wird, setzt voraus, dass keine fehlenden Werte im Datensatz existieren. Das führt dazu, dass im Folgenden nur jeweils Gruppen von Variablen zusammen untersucht werden.
Durch inhaltliche Überlegungen wird vermutet, dass sich unter den Teilnehmern der Umfrage drei Gruppen von Nutzern erkennen lassen:

- „Der Intensiv-Nutzer“: Lernt und arbeitet viel in den Universitätsbibliotheken, nutzt viele der angebotenen Services, verwendet die dort vorhandene Ausstattung.
- „Der phasenweise Nutzer“: Nutzt die Universitätsbibliothek zeitweise vermehrt, z.B. zur Anfertigung einer Abschlussarbeit. Er lernt vermehrt zu Hause, nutzt aber trotzdem einige Services der UB.
- „Der Minimalnutzer“: Entleiht nur gelegentlich Bücher, nutzt nur wenige Services.

4.1. Lernort

Es wird also zunächst eine Clusteranalyse mit drei Gruppen unter Verwendung der neun Variablen zum Lernort durchgeführt. Da diese Variablen alle ordinalskaliert sind, geschieht die Distanzberechnung wie in Abschnitt 2.3.2 beschrieben. Als Zentroid wird der jeweilige Median verwendet. Die Zentroide der drei resultierenden Cluster sind in Grafik 4.1 dargestellt. Dabei entspricht 1.0 der Antwort „fast immer“, 0.0 der Antwort nie. Die roten Punkte markieren dabei die Zentroide der Gesamtheit der Teilnehmer.
Cluster 1 enthält mit 242 Teilnehmern 15% der Teilnehmer. Cluster 2 und Cluster 3 enthalten beide mit 666 Teilnehmern jeweils 42% der Teilnehmer. Teilnehmer aus Cluster 1 lernen häufiger in der Münchener Stadtbibliothek und in den Bibliotheken der TU als die Gesamtheit der Teilnehmer. Der Zentroid von Cluster 2 unterscheidet sich vom Zentroid des gesamten Datensatzes durch eine geringere Nutzung der Fachbibliotheken, der Zentrale der Universitätsbibliothek, der Studentenbibliothek und der Staatsbibliothek. Es liegt die Vermutung nahe, dass sich in diesem Cluster eher die Minimalnutzer befinden. Teilnehmer aus Cluster 3 lernen weniger zu Hause, ansonsten entspricht der Zentroid von Cluster 3 dem Zentroid der Gesamtheit der Teilnehmer.

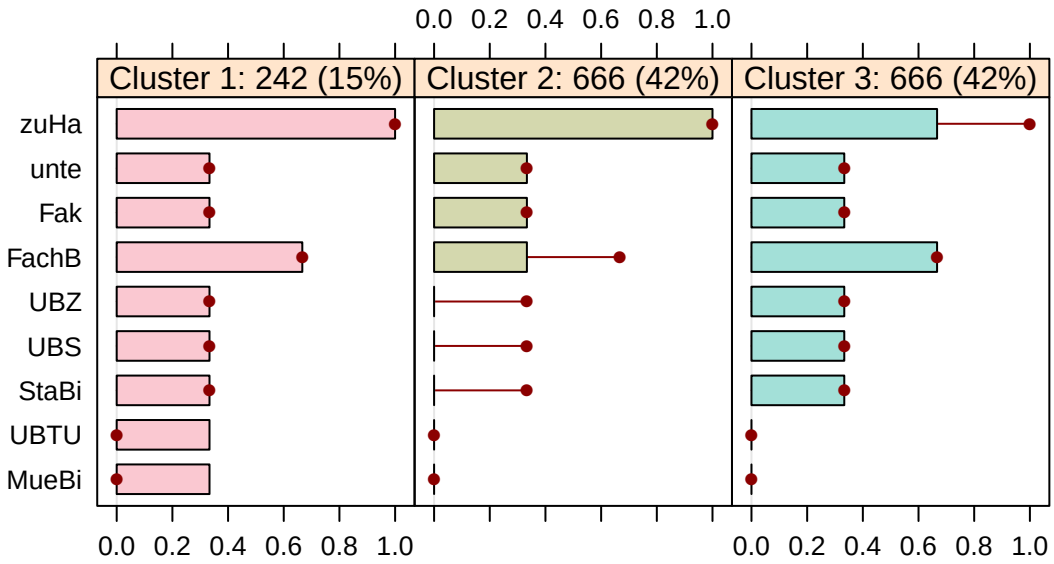

Abbildung 4.1.: Clusterung der Lernorte mit drei Gruppen

Eine erneute Clusterung mit anderen zufälligen Startpositionen führt zu einem nur

leicht unterschiedlichen Ergebnis. Es wurde jedoch wieder ein Cluster mit Teilnehmern gefunden, dessen Zentroid identisch mit dem oben beschriebenen Zentroid von Cluster 2 ist. Eine eindeutige Übereinstimmung mit den zuvor vermuteten Nutzerprofilen kann nicht festgestellt werden. Der Nachbarschaftsgraph (Grafik 4.2) dieser Analyse zeigt eine starke Verbindung zwischen Cluster 1 und Cluster 3. Der Shadow-Plot für diese Clusterung zeigt ebenfalls, dass Cluster 2 am besten separiert ist. Es zeigt sich hier, wie auch später in Abschnitt 4.2, dass die Gruppe der Minimalnutzer am besten separiert ist.

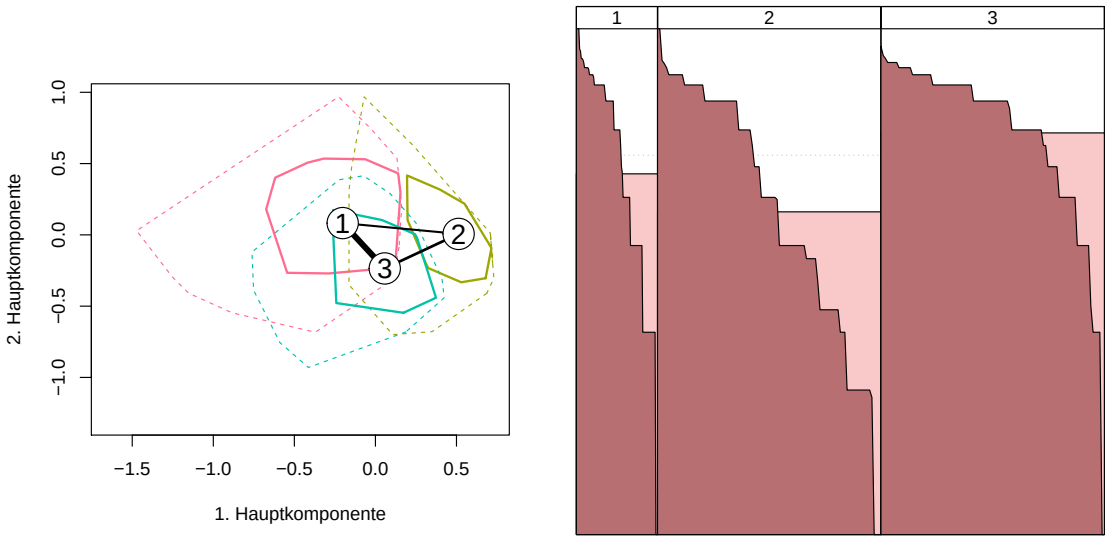

Abbildung 4.2.: Nachbarschaftsgraph und Shadowplot für Clusterung der Lernorte

Der Grund für die Umfrage war der Wunsch der UB, zu erfahren, ob und was sie verändern könnte oder sollte, um den Bedürfnissen der Studierenden noch besser gerecht zu werden. Da den Variablen zu den Lernorten in diesem Zusammenhang nur ein geringes Gewicht beigemessen werden kann, wird auf eine Analyse hinsichtlich der Hintergrundvariablen verzichtet.

4.2. Services

Anschließend werden die genutzten Services betrachtet. Da diese Variablen alle asymmetrisch binär zu interpretieren sind, wird als Distanzmaß die Jaccard-Distanz (Abschnitt 2.3.1) verwendet. Die Zentroidberechnung erfolgt wie in Abschnitt 2.4 beschrieben. Es wurden mehrere zufällige Startpartitionen verwendet und die beste Partitionierung behalten. Die Zentroide der drei resultierenden Cluster sind in Grafik 4.3 dargestellt. Cluster 1 enthält mit mehr als 1000 Beobachtungen ungefähr $\frac{2}{3}$ aller Beobachtungen. In diesem Cluster werden alle Services überdurchschnittlich genutzt. In Cluster 2 werden alle Services unterdurschnittlich stark genutzt. Lediglich OPAC und Ausleihe werden von mehr als 50% der Umfrageteilnehmer genutzt. Cluster 3 ist Cluster 2 sowohl in der Größe als auch darin ähnlich, dass auch hier die meisten Services unterdurchschnittlich viel genutzt werden. Ein Unterschied zu Cluster 2 zeigt sich darin, dass hier Führungen, Infomaterial und die Homepage überdurchschnittlich oft genutzt werden. In allen 3 Clustern wurde der OPAC, die Präsenznutzung und die Ausleihe, wie in der Gesamtheit aller Umfrageteilnehmer sehr stark genutzt. Die Neuerwerbungsliste wurde dagegen in jedem Cluster sowie in der Gesamtheit der Umfrageteilnehmer sehr gering genutzt. Es wäre also zu überlegen, ob ein Ausschluss dieser Services der Übersichtlichkeit dienlich wäre.

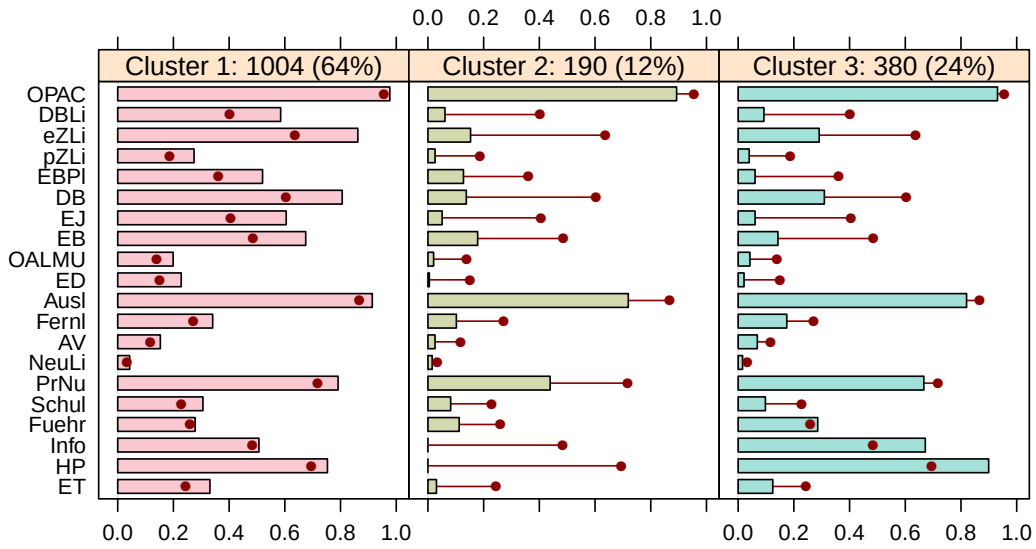

Abbildung 4.3.: Clusterung der Services mit drei Gruppen

Da die Variablen, die die Nutzung der UB-Services wiedergeben, am interessantesten erscheinen, werden im folgenden die Hintergrundvariablen Alter, Geschlecht, Hauptfachrichtung, Semesterzahl und Zufriedenheit untersucht.

Alter

Die deskriptive Analyse der Cluster bezüglich der Altersverteilung innerhalb der Cluster, durch eine Fünf-Punkte-Zusammenfassung (Tabelle 4.1) ergibt Hinweise auf strukturelle Unterschiede zwischen den Clustern.

Cluster	Min.	1st Qu.	Median	3rd Qu.	Max.
1	18.00	22.00	24.00	26.00	75.00
2	18.00	21.00	22.00	24.00	72.00
3	18.00	21.00	23.00	24.00	69.00

Tabelle 4.1.: Fünf-Punkte-Zusammenfassung des Alters für Clusterung nach Services

Es werden deshalb folgende Hypothesen getestet:

- H_0: Die Verteilung der Variable Alter unterscheidet sich hinsichtlich der Lageparameter nicht in den einzelnen Clustern.
- H_1: In mindestens einem Cluster unterscheidet sich die Verteilung der Variable Alter von einem anderen Cluster hinsichtlich der Lageparameter.

Als Annahmen werden getroffen, dass das Alter einer Person unabhängig und innerhalb der Cluster identisch verteilt ist. Die Hypothesen sollen mit einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ getestet werden. Als Teststatistik wird die Kruskal-Wallis Teststatistik wie in Duller (2008, S. 215) verwendet.
Für genügend große Stichprobenumfänge kann als Referenzwert für die Kruskal-Wallis Teststatistik das $(1 - \alpha)$-Quantil der χ^2-Verteilung mit zwei Freiheitsgraden verwendet werden.
Als Wert der Kruskal-Wallis-Teststatistik wird 83.37 errechnet, und ist damit größer als das entsprechende Quantil der χ^2-Verteilung ($\chi^2_{(2;0.95)} = 5.99$). Die Nullhypothese muss somit abgelehnt werden, die Alternativhypothese angenommen werden. Das bedeutet, in mindestens einem Cluster unterscheidet sich die Verteilung der Variable Alter hinsichtlich der Lageparameter von der Verteilung in anderen Clustern.

Geschlecht

Die Relation der Geschlechter innerhalb der jeweiligen Cluster (Grafik 4.4) beträgt in Cluster 3 ungefähr 25% Männer, in Cluster 1 und Cluster 2 dagegen jeweils etwa 37% Männer.
Um einen χ^2-Unabhängigkeitstest durchführen zu können, wird die Clusterzugehörigkeit als eine Variable aufgefasst. Die mit einem Irrtumsfehler $\alpha = 0.05$ zu testenden Hypothesen lauten:

- H_0: Die Variablen Clusterzugehörigkeit und Geschlecht sind stochastisch unabhängig.
- H_1: Die Variablen Clusterzugehörigkeit und Geschlecht sind stochastisch abhängig.

Die Voraussetzungen für die Anwendbarkeit dieses Tests, wie in Abschnitt 2.6.1 vorgestellt, sind in diesem Fall erfüllt. Als Wert der χ^2-Teststatistik ergibt sich 10.71 und ist damit größer als $\chi^2_{(2;0.95)} = 5.99$. Die Nullhypothese wird damit verworfen, es besteht ein signifikanter Zusammenhang zwischen der Clusterzugehörigkeit und dem Geschlecht.

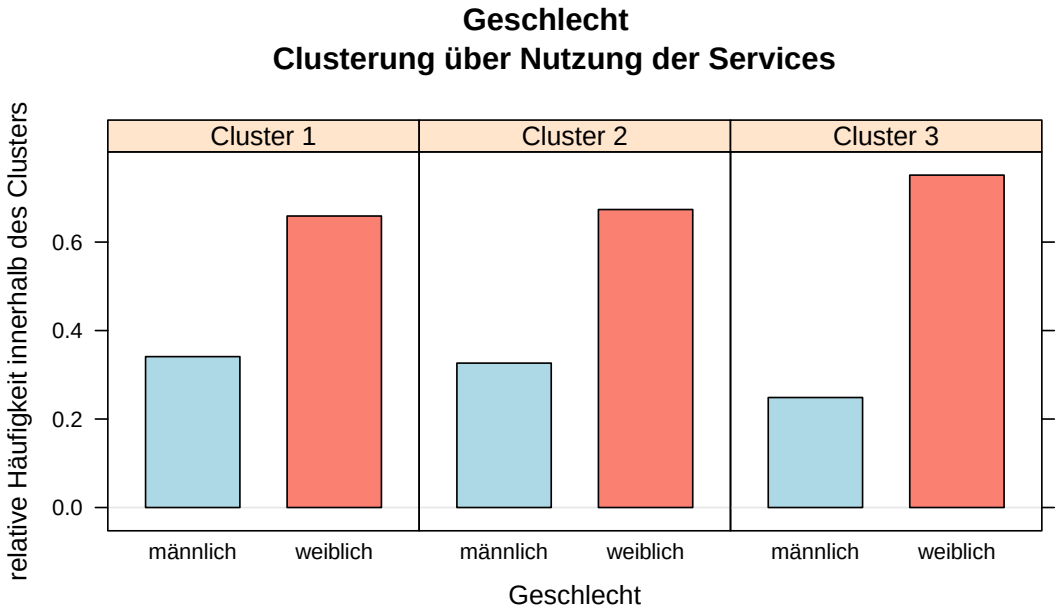

Abbildung 4.4.: Geschlecht innerhalb der Cluster

Hauptfachrichtung

Bei der Analyse der Hauptfachrichtungen zeigt sich, dass in Cluster 1 und 3 eine ähnliche Aufteilung der Fachrichtungen, wie in der Gesamtheit der Umfrageteilnehmer existiert (siehe Eifler u. a. 2011, S. 7). Cluster 2 weist dagegen einen höheren Anteil an Studierenden mit Fachrichtung Mathematik und Naturwissenschaften, sowie Medizin und Gesundheitswissenschaften auf. Dies ist in Grafik 4.5 dargestellt. Auch hier wird wieder die Hypothese getestet, ob die Variablen Clusterzugehörigkeit und Studienfach stochastisch unabhängig sind. Als Wert der Teststatistik ergibt sich 55.45. Dies ist größer, als das entsprechende Quantil der χ^2-Verteilung ($\chi^2_{(6;0.95)} = 12.59$), die als Referenzwert für einen Irrtumsfehler $\alpha = 0.05$ verwendet wird. Die Nullhypothese, dass Studienfach und Clusterzugehörigkeit stochastisch unabhängig voneinander sind, wird verworfen.

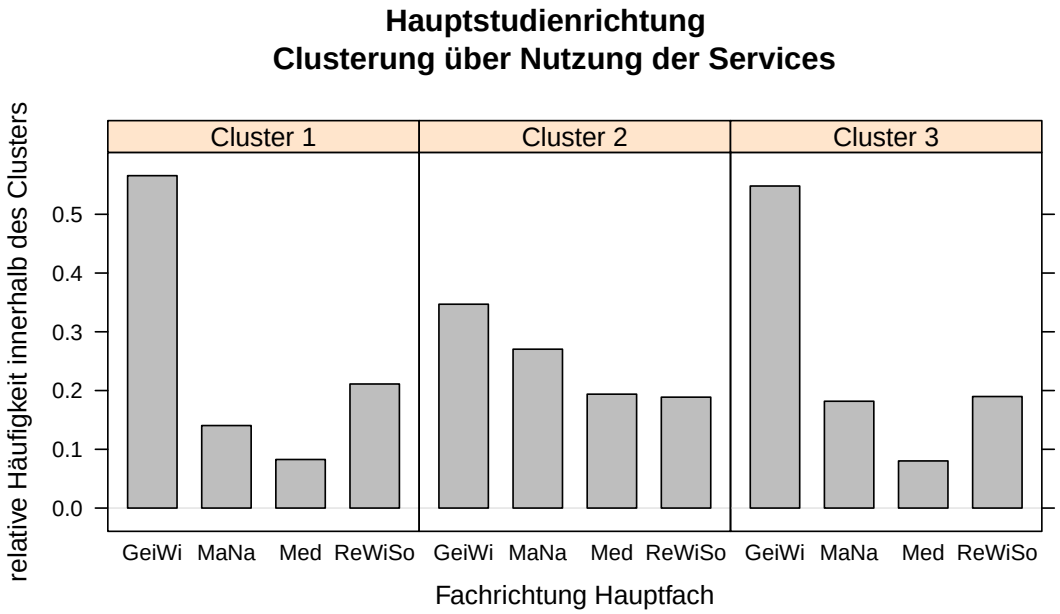

Abbildung 4.5.: Hauptfachrichtungen innerhalb der Cluster

Semesterzahl

Eine deskriptive Analyse der Cluster bezüglich der Semesterzahl (Grafik 4.6) zeigt, dass sich in Cluster 3 mit über 20% Erstsemestern viele Studienanfänger befinden. In Cluster 1 dagegen befinden sich etwa 5% Erstemestern, der Modus liegt bei neun Semestern. Dies könnte ein Hinweis darauf sein, dass Studienanfänger vor allem Führungen, Homepage und Infomaterial nutzen, während weiter fortgeschrittene Studierende alle Services in größerem Umfang beanspruchen.

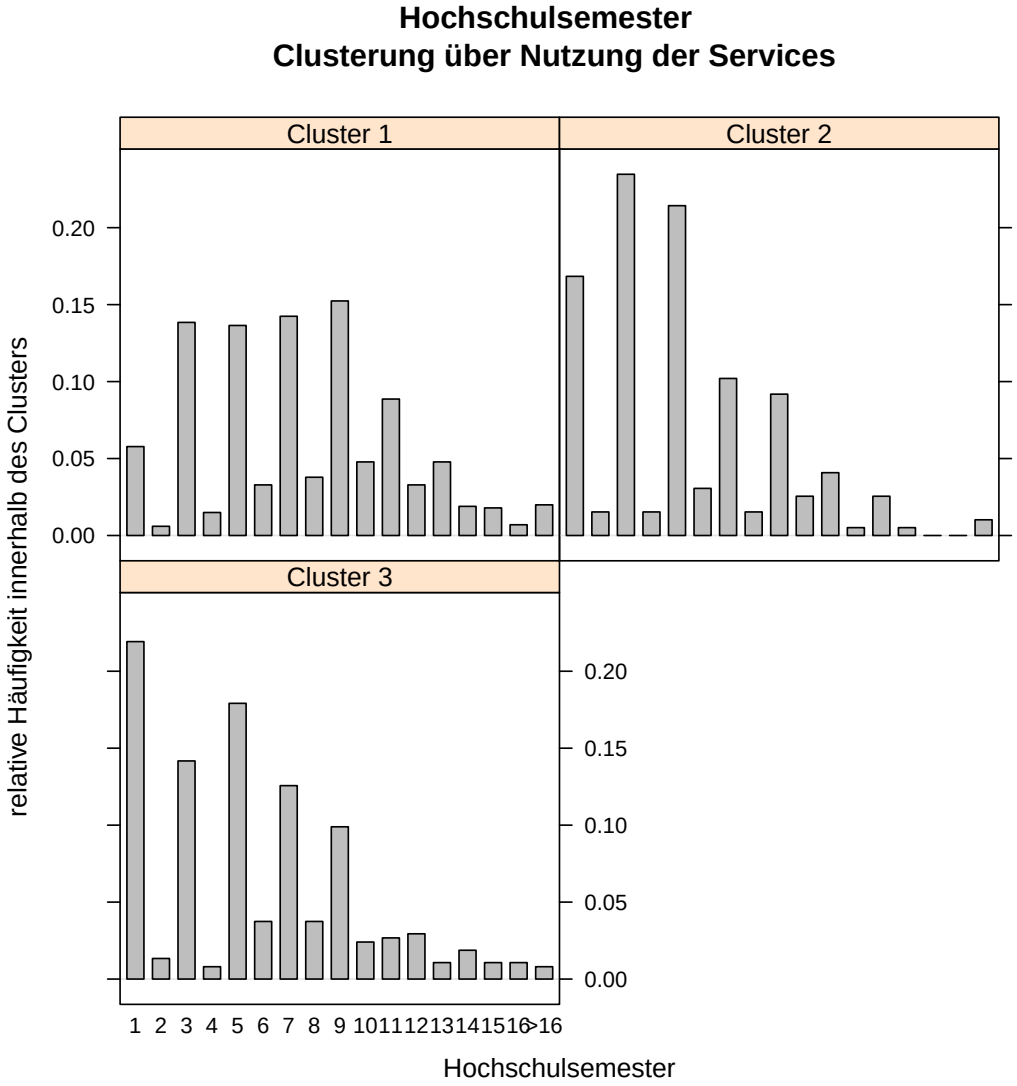

Abbildung 4.6.: Semesterzahl innerhalb der Cluster

Zufriedenheit

Als nächstes wird untersucht, ob sich ein Zusammenhang zwischen der Zugehörigkeit zu einem Cluster und der Zufriedenheit in Bezug auf den persönlichen Support und den Informationsfluss herstellen lässt. Dazu wird jeweils der Mittelwert der drei Variablen, die in Abschnitt 3.3 beschrieben wurden, verwendet. Die Grafik 4.7 zeigt in Cluster 1 und Cluster 3 eine ähnliche Aufteilung. In Cluster 2 ist die durchschnittliche Anzahl an neutralen Antworten am größten. Dies zeigt ebenso wie der Zentroid von Cluster 2, dass sich hier möglicherweise Studienteilnehmer mit geringerem Interesse an der UB befinden.

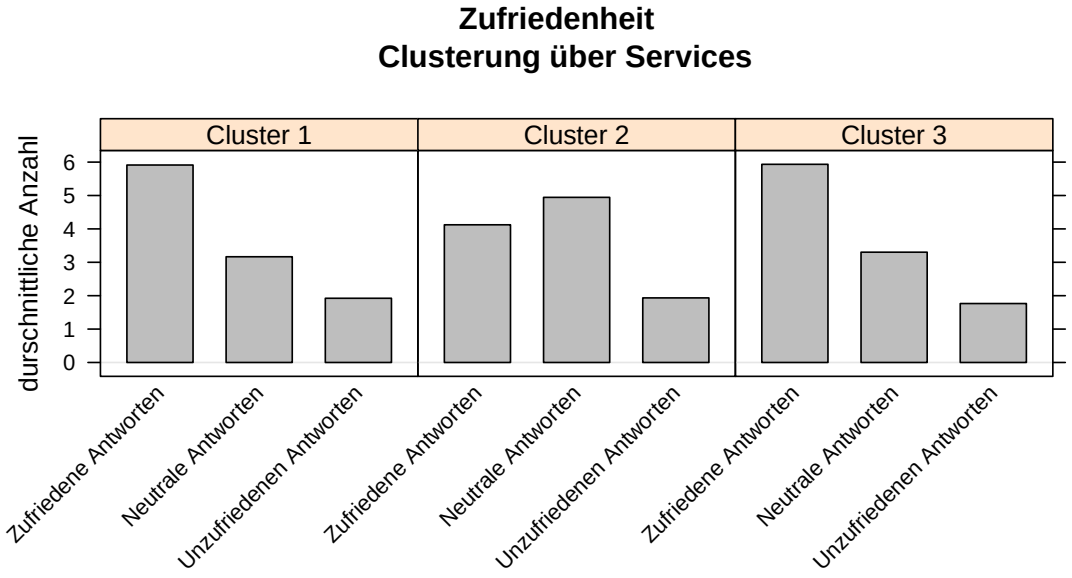

Abbildung 4.7.: Zufriedenheit innerhalb der Cluster

Versucht man, die Cluster den angenommenen Nutzerprofilen zuzuordnen, so ergibt sich, dass Cluster 1 den „Intensiv-Nutzern“ und den „phasenweise-Nutzer“ entsprechen könnte, während Cluster 2 eher den Minimalnutzern entspricht. Die Nutzer in Cluster 3 haben ein starkes Bedürfnis nach Informationsangeboten. In diesem Cluster befinden sich viele Studienanfänger. Sie nutzen die Services der UB (noch) nicht so intensiv. Es kann also sein, dass die zuvor überlegten Nutzerprofile korrigiert werden müssen. Diese Vermutung wird ebenfalls durch den Nachbarschaftsgraphen (Grafik 4.8) für diese Clusteranalyse unterstützt, denn er zeigt eine starke Verbindung zwischen den Clustern 1 und 3. Dies könnte ein Hinweis darauf sein, dass sich vor allem 2 Nutzerprofile in den Daten wiederfinden lassen.

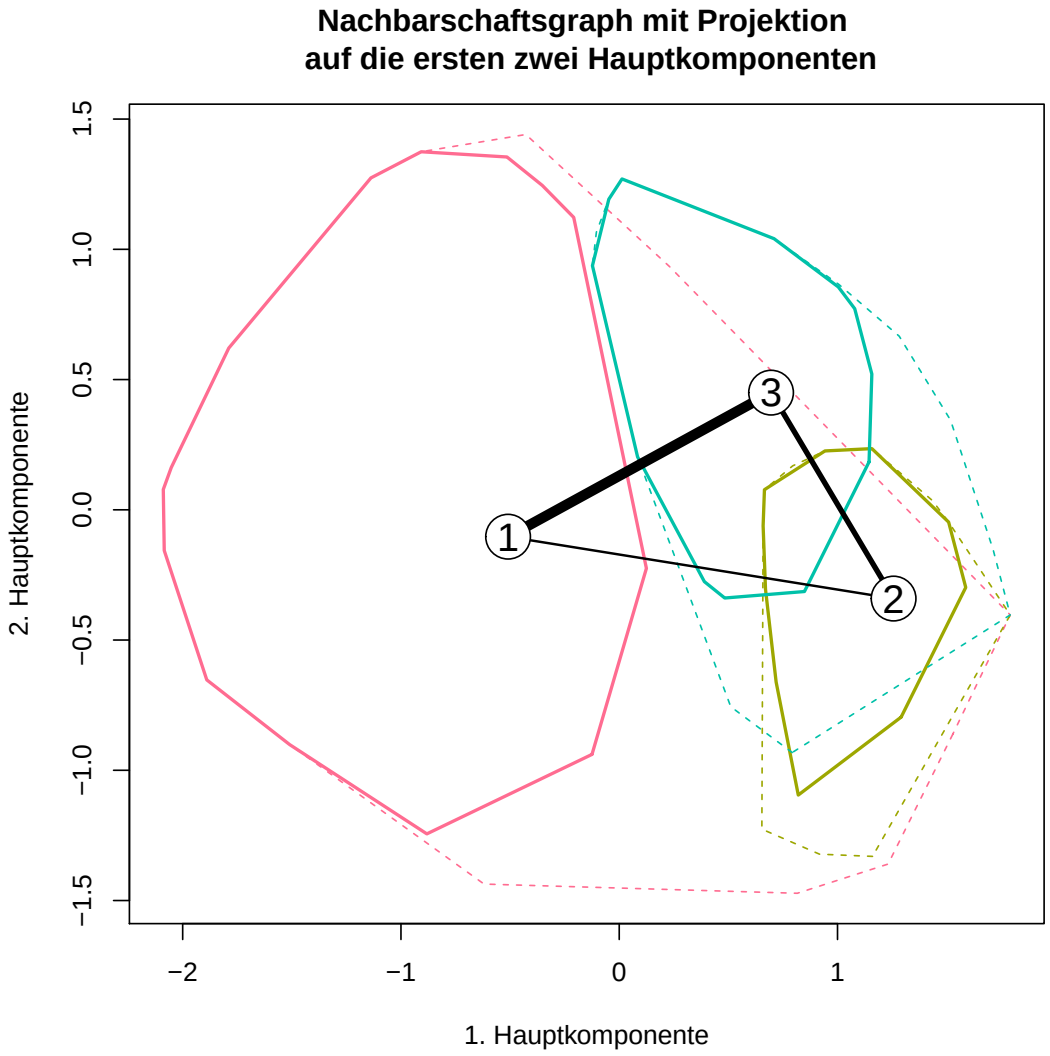

Abbildung 4.8.: Nachbarschaftsgraph für Clusterung der Services

4.3. Lernzeiten

Als nächstes sollen die Lernzeiten untersucht werden. Dazu werden die insgesamt 30 binären Variablen zur Frage: „Zu welchen Zeiten lernen Sie?“ betrachtet. In diesem Fall sagt eine Eins wesentlich mehr über die Ähnlichkeit zweier Personen aus, als eine Null. Eins bedeutet, zwei Personen lernen zur gleichen Zeit. Null heißt, sie könnten entweder zu dieser Zeit nicht lernen, oder aber die Frage einfach übersprungen haben. Deswegen

werden die Variablen zur Lernzeit als asymmetrisch betrachtet und zur Clusterung die Jaccard-Distanz mit erwartungsbasierten Zentroiden verwendet.
Da hier kein intuitiver Zusammenhang mit den Nutzerprofilen hergestellt werden kann, werden, wie in Abschnitt 2.4 vorgeschlagen, zunächst verschiedene Clusteranzahlen ausprobiert und mithife eines Scree-Plots (Grafik 4.9) verglichen. Für jede Clusteranzahl wurden wieder mehrere zufällige Startpartitionen verwendet und die beste Partitionierung behalten.

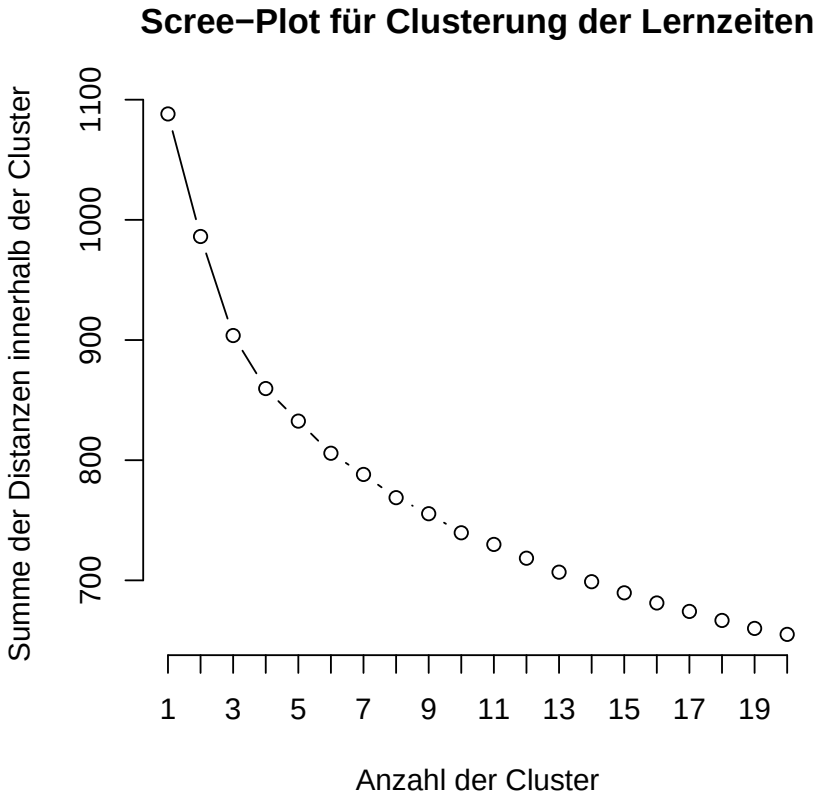

Abbildung 4.9.: Scree-Plot für 1-20 Cluster der Lernzeiten

Der Scree-Plot zeigt keinen „klaren Knick“ und gibt somit keinen Aufschluss über eine „wahre“ Anzahl an Clustern. Eine anschließende Analyse der Shadow-Plots und Nachbarschaftsgraphen für die Clusterlösungen mit zwei bis neun Zentroiden führte zu einer Entscheidung zwischen drei und fünf Clustern. Es soll allerdings auf Grund der Übersichtlichkeit und Interpretierbarkeit an dieser Stelle nur auf die Lösung mit drei

Clustern eingegangen werden.
Die Grafik 4.10 zeigt den Nachbarschaftsgraphen und den Shadow-Plot für diese Partitionierung.

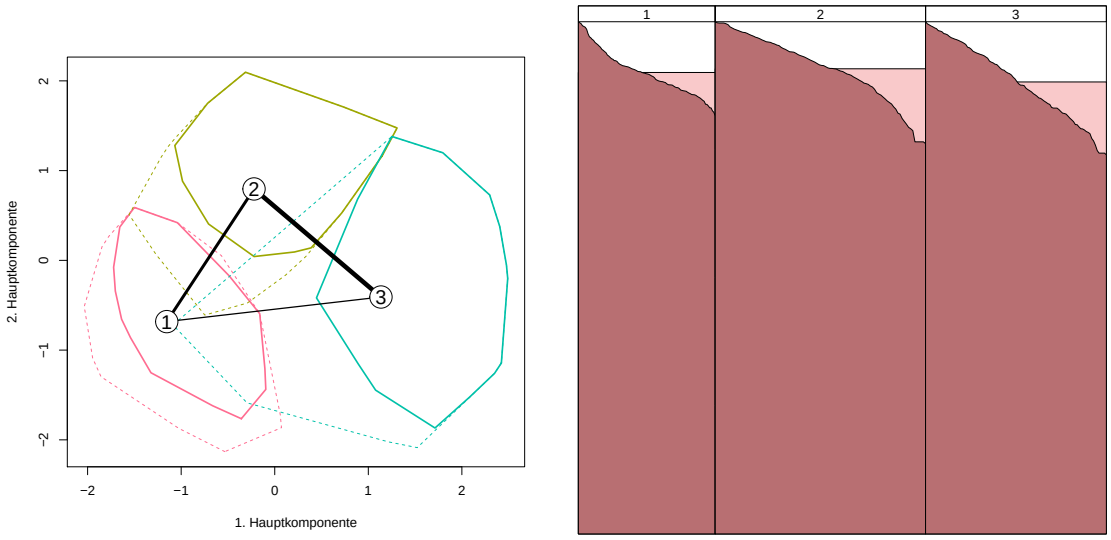

Abbildung 4.10.: Nachbarschaftsgraph und Shadow-Plot für 3 Clusterlösung der Lernzeiten

Die Shadow-Werte bewegen sich in allen drei Clustern zwischen 0.7 und 1. Der Nachbarschaftsgraph zeigt die stärkste Verbindung zwischen den Clustern 2 und 3, aber alle Cluster haben Verbindungen untereinander. Dies sind starke Indizien für eine künstliche Partitionierung. Es kann aber trotzdem Sinn ergeben, diese Gruppen auf Unterschiede zu untersuchen, um Angebote der UB, wie z.B. Öffnungszeiten besser anpassen zu können. (vgl. Leisch 2010, S. 462)
Die folgende Tabelle 4.2 zeigt die Clustergrößen. Dabei ist auffällig, dass anders als in Abschnitt 4.2 die Clustergrößen jetzt ähnlicher sind. Eine Darstellung der Zentroide,

Cluster	absolute Größe	relative Größe
1	408	26%
2	627	40%
3	539	36%

Tabelle 4.2.: Tabelle der Clustergrößen für Lernzeiten

wie bei den Analysen der Nutzerprofile wird an dieser Stelle aufgrund der hohen Variablenzahl nur zusammengefasst. Die entsprechende Grafik A.1 befindet sich deswegen im

Anhang. Der deutlichste Unterschied besteht zwischen Cluster 1 und den beiden anderen Clustern durch das Lernverhalten in den Ferien. So wurde in Cluster 1 überwiegend „nie in den Ferien gelernt“. Cluster 2 und Cluster 3 unterscheiden sich nicht so stark voneinander, was auch der Nachbarschaftsgraph (Grafik 4.10) zeigt. Der Unterschied liegt in diesem Fall darin, dass Teilnehmer in Cluster 2 vermehrt vormittags lernen, während Teilnehmer in Cluster 3 vermehrt abends und nachts lernen.

Alter

Die Grafik 4.11 zeigt die Boxplots des Alters. Es wird auch hier ein Kruskal-Wallis-Tests durchgeführt. Die Nullhypothese, dass die Verteilungen des Alters in den jeweiligen Clustern sich hinsichtlich ihrer Lageparameter nicht unterscheiden wird wie in Abschnitt 2.6.2 beschrieben durchgeführt. Als Irrtumswahrscheinlichkeit wird $\alpha = 0.05$ angesetzt. Als Wert der Teststatistik ergibt sich 23.04. Der Vergleich mit dem entsprechenden Quantil der χ^2-Verteilung ($\chi^2_{(2;0.95)} = 5.99$) führt zur Ablehnung der Nullhypothese.

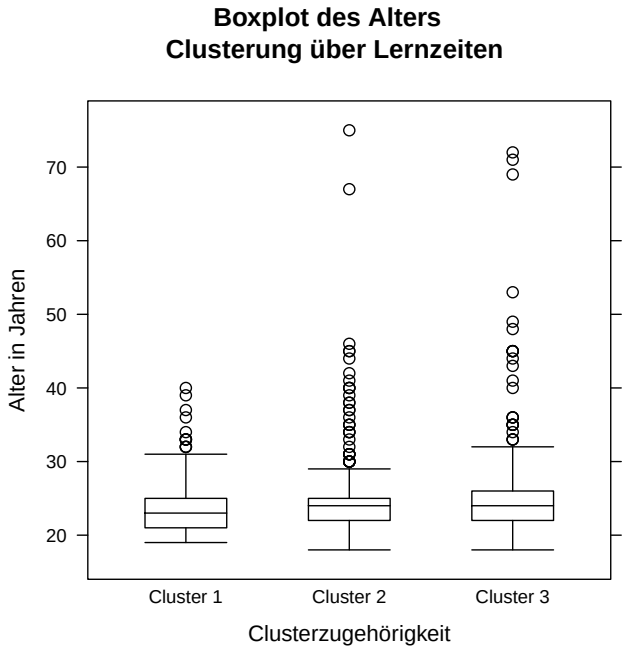

Abbildung 4.11.: Boxplot des Alters für die Lernzeiten

Geschlecht

Die Analyse der relativen Häufigkeiten des Geschlechts innerhalb der Cluster durch Tabelle 4.3 zeigt, dass sich in Cluster 3 mit 37% die meisten Männer befinden. Cluster 1 und 2 sind sich in Bezug auf die Geschlechteranteile jedoch sehr ähnlich.

Cluster	männlich	weiblich
1	0.31	0.69
2	0.28	0.72
3	0.37	0.63

Tabelle 4.3.: Tabelle der relativen Geschlechteranteile für Clusterung nach Lernzeiten

Der höhere Männeranteil in Cluster 3 legt die Vermutung nahe, dass Männer eventuell eher geneigt sind abends zu lernen. Analog zu Abschnit 4.2 wird wieder ein χ^2-Unabhängigkeitstest durchgeführt. Die Nullhypothese lautet: Die Variablen Geschlecht und Clusterzugehörigkeit sind stochastisch unabhängig. Als Irrtumswahrscheinlichkeit wird $\alpha = 0.05$ festgelegt. Als Wert der χ^2-Teststatistik ergibt sich 12.05 und ist damit größer als $\chi^2_{(2;0.95)} = 5.99$. Die Nullhypothese wird damit verworfen, es besteht ein signifikanter Zusammenhang zwischen der Clusterzugehörigkeit und dem Geschlecht.

Hauptfachrichtung

Die Grafik 4.12 zeigt die relativen Häufigkeiten der Studienfächer in allen drei Clustern. Es wird erneut die Nullhypothese getestet, dass Studienfach und Clusterzugehörigkeit stochastisch unabhängig sind. Als Irrtumsfehler wird $\alpha = 0.05$ verwendet. Der Wert der χ^2-Teststatistik beträgt 16.73 und ist damit größer als $\chi^2_{(6,0.95)} = 12.59$. Die Nullhypothese wird damit verworfen. Es gibt einen signifikanten Zusammenhang zwischen der Clusterzugehörigkeit und dem Studienfach. Dieser Zusammenhang scheint jedoch nicht substantiell, da die Aufteilung der relativen Häufigkeiten in den drei Clustern sehr ähnlich ist.

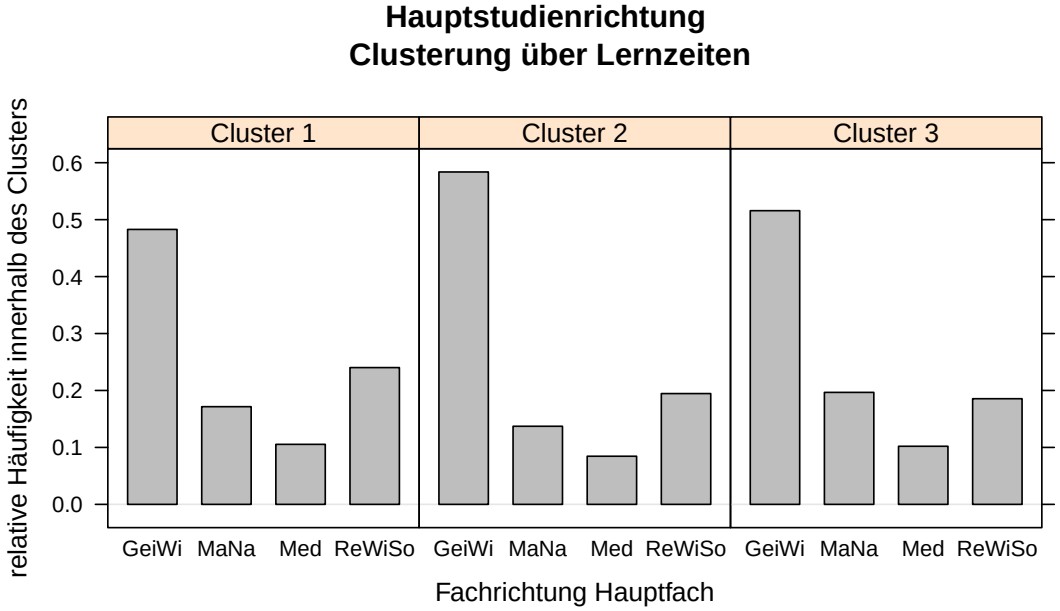

Abbildung 4.12.: Hauptfachrichtung innerhalb der Cluster

Semesterzahl

Die Aufteilung der Semester innerhalb der Cluster (Grafik 4.13) lässt dagegen Unterschiede zwischen den Clustern erkennen. So ist in Cluster 1 der Modus sehr ausgeprägt im dritten Semester, während in den beiden anderen Clustern kein deutlicher Modus existiert. Dafür sind in Cluster 3 mehr Studierende mit mehr als 10 Semestern. In Cluster 2 befindet sich ein Großteil der Studierenden zwischen dem siebten und dem elften Semester.

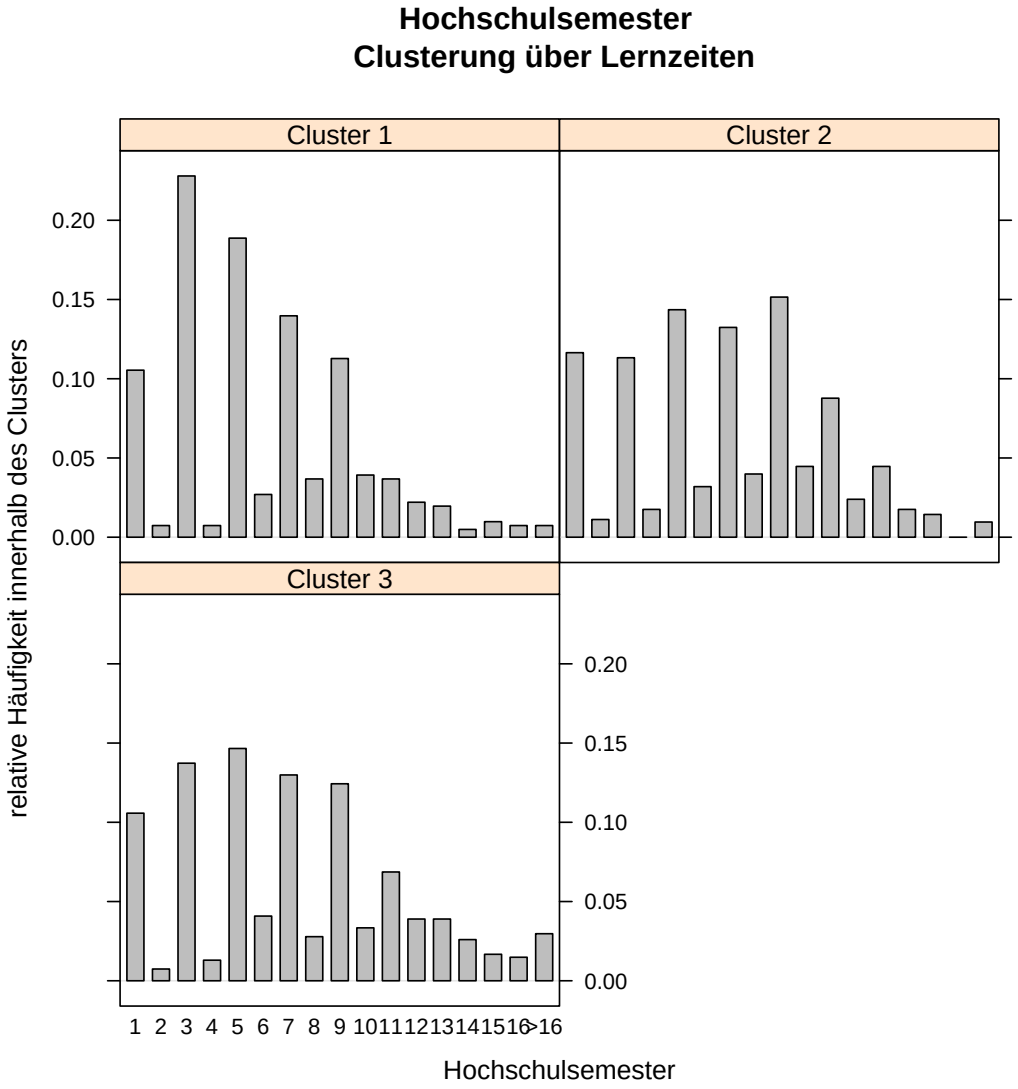

Abbildung 4.13.: Verteilung der Semester innerhalb der Cluster

Das Ergebnis dieser Analysen führt zu folgenden Vermutungen: Studierende in den ersten fünf Semestern lernen seltener in den Semesterferien. Studierende in höheren Semestern lernen vermehrt auch in den Semesterferien. Möglicherweise lernen männliche Studenten eher abends. Seniorenstudenten sind vermehrt bei den „Ferienlernern“ (Cluster 2 und 3) zu finden.

4.4. Ausstattungswünsche und Services

Nach der Analyse der Services im Rahmen der Nutzerprofile (Abschnitt 4.2), wurde vermutet, dass OPAC, Präsenznutzung, Ausleihe und Neuerwerbungsliste nur geringen Einfluss haben. Diese Services sollen deswegen in diesem Abschnitt nicht mehr in die Untersuchungen eingehen. Stattdessen werden vier Variablen aus der Frage zu den Ausstattungswünschen hinzugenommen. Diese sind „Bücher aus der Bibliothek“, „E-Medien“, „Scanner/Kopierer/Drucker“ und „Bibliothekscomputer“. Diese Auswahl wird nicht willkürlich getroffen, sondern beruht einerseits auf dem Wissen, dass diese Variablen eine genügend große Anzahl unterschiedlicher Antworten aufweisen (Eifler u. a. 2011, S. 21). Andererseits besteht hier im weitesten Sinne noch ein Bezug zu den Services der UB. Durch die Möglichkeit, bei den Ausstattungsmerkmalen keine Angaben zu machen, enstehen einige fehlenden Werte. Diese können nicht sinnvoll ersetzt werden und daher müssen entsprechende Beobachtungen für diese Clusteranalyse verworfen werden. Die Berechnungen in diesem Fall finden also mit 1420 Studierenden statt. Da die Variablen der Ausstattungsmerkmale ordinales Skalenniveau haben, die Services jedoch asymmetrisch binär zu verstehen sind, berechnet man die Distanz wie in Abschnitt 2.3.3 beschrieben. Das Verfahren der Zentroidberechnung für Daten mit gemischten Skalenniveaus ist in Abschnitt 2.4 beschrieben. Durch eine vierfache Gewichtung des binären Teils der Distanzberechnung wird verhindert, dass die ordinalen Variablen aufgrund ihrer geringeren Anzahl die Clusterbildung stärker beeinflussen.
Analog zu Abschnitt 4.3 werden auch hier zunächst verschiedene Clusteranzahlen ausprobiert und mithilfe eines Screeplots verglichen.

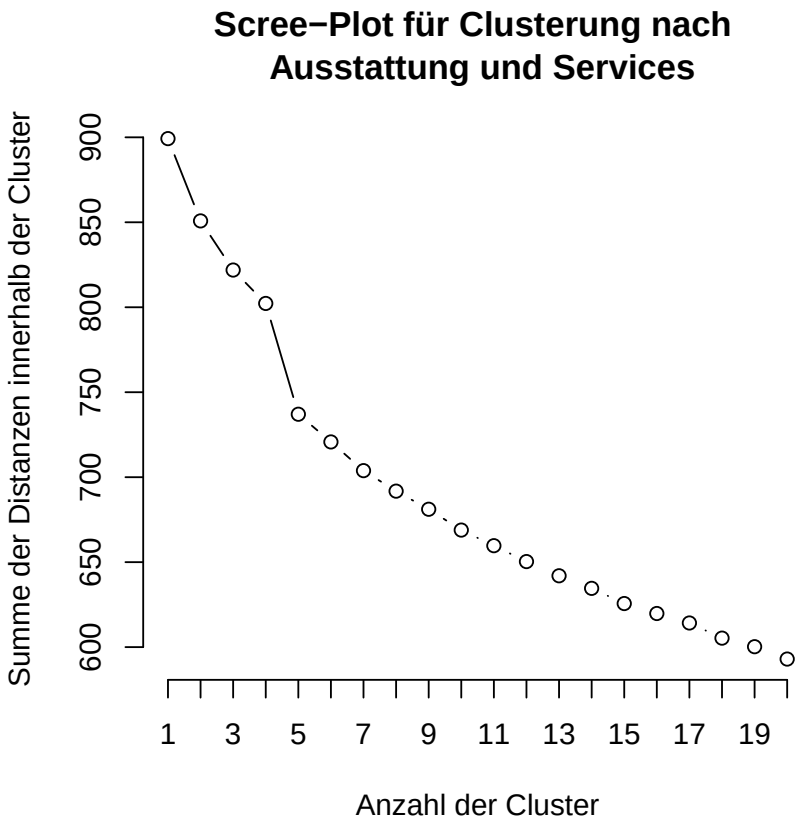

Abbildung 4.14.: Scree-Plot für 1-20 Cluster

In dieser Grafik existiert ein „Knick“ bei einer Clusteranzahl von 5, deswegen wird diese Clusterung im Folgenden genauer betrachtet.
Die Darstellung der Zentroide dieser Clusteranalyse befindet sich im Anhang (Grafik A.2). Auffällig sind zunächst die stark unterschiedlichen Clustergrößen. Die beiden Extreme bilden dabei Cluster 1 mit 65 Studierenden (5%) und Cluster 5 mit 815 Studierenden (57%). In Cluster 2, 3 und 4 befinden sich, mit 186 (13%), 197 (14%) und 158 (11%), in etwa gleich viele Studierende. Um einen besseren Überblick über die Zusammenhänge und die Separiertheit der einzelnen Cluster zu erhalten, wird auch hier der Nachbarschaftsgraph und der Shadow-Plot betrachtet (Grafik 4.15).

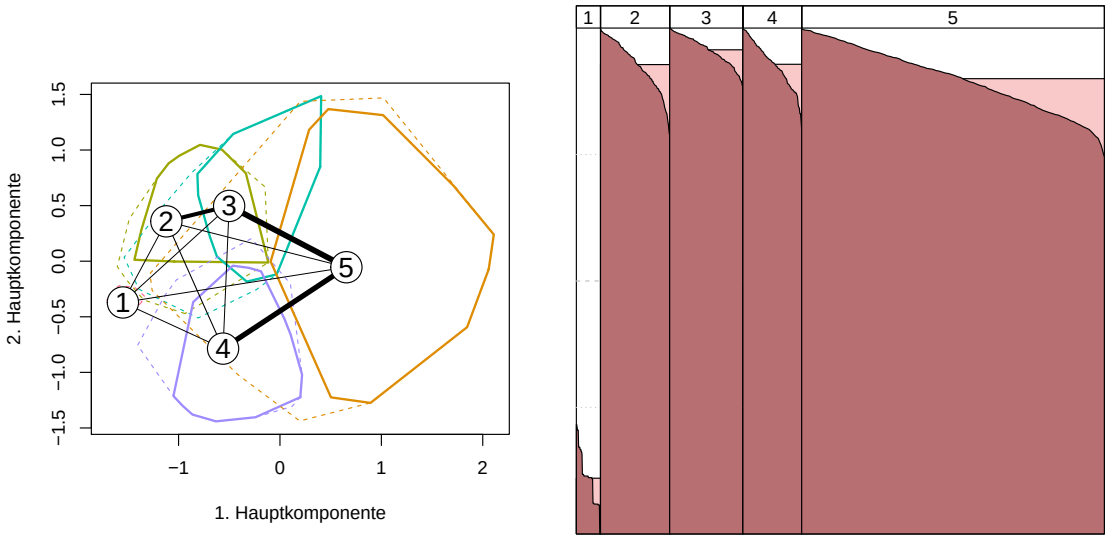

Abbildung 4.15.: Nachbarschaftsgraph und Shadow-Plot für 5 Cluster zur Kombination von Ausstattung und Services

Beide Darstellungsformen zeigen, dass Cluster 1 am besten von den anderen Clustern separiert ist. Cluster 2, Cluster 3, Cluster 5 und Cluster 4 zeigen eine starke Verbindung in dieser Reihenfolge.
Cluster 1 zeichnet sich dadurch aus, dass die Teilnehmer dieser Gruppe, die Services der UB (ausgenommen OPAC, Ausleihe, Präsenznutzung und Neuerwerbungsliste) nicht nutzen und auch bei den Ausstattungswünschen mit Ausnahme der E-Medien anspruchsloser als die Gesamtheit der Teilnehmer sind. Teilnehmer in Cluster 2 unterscheiden sich von Cluster 1 vor allem in 3 Variablen: Sie erachten Scanner/Kopierer/Drucker für wichtig und nutzen überdurchschnittlich die Services Infomaterial und Führungen. Cluster 3 zeichnet sich durch eine höhere Nutzung der Datenbanken, sowie ein stärkeres Interesse am Ausstattungsmerkmal Bibliothekscomputer aus. In Cluster 4 werden die Services E-Book-Plattform und E-Books überdurchschnittlich oft genutzt, die anderen Services der UB jedoch weniger als vom Durchschnitt der Teilnehmer. In Cluster 5, dem größten Cluster, befinden sich die Teilnehmer, die alle Services der UB überdurchschnittlich häufig nutzen. Diese Teilnehmer erachten alle Ausstattungsmerkmale mit Ausnahme des Bibliothekscomputers als wichtig. Die Untersuchung der Cluster bezüglich der Hintergrundvariablen Alter, Geschlecht, Hauptfachrichtung, Semesterzahl und Zufriedenheit würde analog zu den Abschnitten 4.2 und 4.3 erfolgen und wird deshalb hier nicht dargestellt.

5. Schluss

5.1. Zusammenfassung der Ergebnisse

Folgende Probleme traten bei der Analyse der Daten auf: Durch die Zulassung von Mehrfachnennung bei einigen Fragen war es möglich, dass inkonsistente Antworten entstanden. Beobachtungen, bei denen Inkonsistenzen auftraten, wurden teilweise nicht berücksichtigt. Durch viele optionale und konditionierte Fragen enthielt der Datensatz bei vielen Variablen fehlende Werte. Diesem Problem wurde durch eine gezielte Selektion von Variablen mit wenigen fehlenden Werten begegnet. Durch die hohe Anzahl an Variablen bei einigen untersuchten Fragen ergaben sich Schwierigkeiten bei der Darstellung und Bewertung der Analysen. Es sollte bereits bei der Datenerhebung genau darauf geachtet werden, welche Hypothesen getestet werden sollen. Außerdem wäre eine einheitliche Anzahl an Antwortkategorien wünschenswert. Des weiteren sollte bereits bei der Fragebogenerstellung darauf geachtet werden, eine möglichst repräsentative Umfrage durchzuführen, um anschließend gesicherte Aussagen über die Studierenden der Ludwig-Maximilian-Universität treffen zu können. Es ist zu vermuten, dass vor allem Studierende, die die UB und deren Angebote häufiger nutzen, motiviert waren, den Fragebogen auszufüllen.
Trotz der genannten Probleme waren Untersuchungen möglich. Es wurden folgende Fragengruppen gemeinsam untersucht:

1. Lernort
2. Services
3. Lernzeit
4. Ausstattungswünsche gemeinsam mit Services

Die K-Zentroid Clusteranalyse stellte sich dabei als effizientes Verfahren heraus, jedoch ergaben sich Schwierigkeiten bei der Interpretation, wenn viele Variablen gemeinsam untersucht wurden, beziehungsweise eine hohe Clusteranzahl zu untersuchen war. Die

entstandenen Cluster wurden bezüglich ihrer Zentroide, sowie der Hintergrundvariablen Alter, Geschlecht, Fachrichtung und Semesterzahl untersucht. Für Punkt 2 wurde zusätzlich die zusammengesetzten Variable zur Zufriedenheit betrachtet.
Die Untersuchungen zu Punkt eins ergaben keine inhaltlich sinnvoll interpretierbaren Ergebnisse. Zu Punkt zwei stellte sich heraus, dass die Vorstellung von drei Nutzerprofilen (durch Überlegung) in diesem Datensatz nur teilweise wiedergefunden werden konnte. Erkennbar waren vor allem, eine Gruppe von Intensiv-, und eine Gruppe von Minimalnutzern. Zu Punkt drei stellte sich heraus, dass Umfrageteilnehmer in höheren Semestern sowie ältere Studierende eher in den Semesterferien lernen, als Studierende in niedrigeren Semestern. Studierende, die verstärkt abends und nachts lernen, wiesen im Vergleich mit der Gesamtheit der Umfrageteilnehmer einen höheren Anteil männlicher Teilnehmer auf. Unter Punkt vier konnte erneut eine Aufteilung der Umfrageteilnehmer in Intensiv- und Minimalnutzer gefunden werden.

5.2. Ausblick

Es wurden in dieser Arbeit aufgrund der begrenzten Arbeitszeit nur einige der möglichen Clusterlösungen untersucht. Weitere Analysen könnten aus Untersuchungen noch weiterer Variablenkombinationen bestehen. Ebenfalls interessant wäre, ob durch eine unterschiedliche Gewichtung verschiedener Variablen oder der Verwendung anderer Distanzmaße bei der Clusteranalyse weitere Strukturen innerhalb der Daten gefunden werden können. Schließlich sei erwähnt, dass der in dieser Arbeit zur Clusterbildung verwendete Algorithmus nicht der einzige zur Analyse dieses Datensatz denkbare Algorithmus ist. Es wäre also zum Beispiel interessant, ob durch die Verwendung von Medoiden statt Zentroiden ähnliche Ergebnisse erzielt werden können.

Literaturverzeichnis

Backhaus u. a. 2008

Backhaus, Klaus ; Erichson, Bernd ; Plinke, Wulff ; Weiber, Rolf: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. Springer, 2008 (Springer-Lehrbuch Series). – ISBN 978–3–540–85044–1

Dahl 2009

Dahl, David B.: xtable: Export tables to LaTeX or HTML, 2009. http://CRAN.R-project.org/package=xtable. – R package version 1.5-6

Duller 2008

Duller, Christine: Einführung in die nichtparametrische Statistik mit SAS und R: Ein anwendungsorientiertes Lehr-und Arbeitsbuch. Physica-Verlag, 2008. – ISBN 978–3–7908–2059–1

Eifler u. a. 2011

Eifler, Fabian ; Pietsch, Robert ; Schierholz, Malte: Bericht zum Statistischen Praktikum: Studierendenbefragung der Universitätsbibliothek. http://www.cip.ifi.lmu.de/~pietsch/bericht.pdf. Version: 2011

Everitt 1974

Everitt, Brian: Cluster Analysis. Heinemann Educational for the Social Science Research Council, 1974 (Reviews of Current Research). – ISBN 0435822977

Fahrmeir u. a. 1996

Fahrmeir, Ludwig ; Hamerle, Alfred ; Tutz, Gerhard: Multivariate statistische Verfahren. Walter de Gruyter, 1996. – ISBN 3–11–013806–9

Hothorn u. a. 2010

Hothorn, Torsten ; Leisch, Friedrich ; Zeileis, Achim: modeltools: Tools and Classes for Statistical Models, 2010. http://CRAN.R-project.org/package=modeltools. – R package version 0.2-17

Kaufman u. Rousseeuw 2005

Kaufman, Leonard ; Rousseeuw, Peter J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 2005 (Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics). – ISBN 978–0–471–76578–6

Leisch 2006

Leisch, Friedrich: A Toolbox for K-Centroids Cluster Analysis. In: Computational Statistics and Data Analysis 51 (2006), Nr. 2, S. 526–544. http://dx.doi.org/10.1016/j.csda.2005.10.006. – DOI 10.1016/j.csda.2005.10.006. – ISSN 0167–9473

Leisch 2010

Leisch, Friedrich: Neighborhood Graphs, Stripes and Shadow Plots for Cluster Visualization. In: Statistics and Computing 20 (2010), S. 457–469. http://dx.doi.org/10.1007/s11222-009-9137-8. – DOI 10.1007/s11222–009–9137–8. – ISSN 0960–3174

Martinetz u. Schulten 1994

Martinetz, Thomas ; Schulten, Klaus: Topology Representing Networks. In: Neural Networks 7 (1994), Nr. 3, S. 507–522. http://dx.doi.org/10.1016/0893-6080(94)90109-0. – DOI 10.1016/0893–6080(94)90109–0. – ISSN 0893–6080

R Development Core Team 2010

R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2010. http://www.R-project.org. – ISBN 3-900051-07-0

Rousseeuw 1987

Rousseeuw, Peter J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. In: Journal of Computational and Applied Mathematics 20 (1987), S. 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7. – DOI 10.1016/0377–0427(87)90125–7. – ISSN 0377–0427

Sarkar 2008

Sarkar, Deepayan: Lattice: Multivariate Data Visualization with R. New York : Springer, 2008 http://lmdvr.r-forge.r-project.org. – ISBN 978-0-387-75968-5

Wedel u. Kamakura 1998

Wedel, Michel ; Kamakura, Wagner A.: Market Segmentation: Conceptual and

Methodological Foundations. Kluwer Academic, 1998 (International Series in Quantitative Marketing). – ISBN 9780792380719

Wikipedia 2010

WIKIPEDIA: Konstitutionspsychologie – Wikipedia, Die freie Enzyklopâdie. http://de.wikipedia.org/w/index.php?title=Konstitutionspsychologie&oldid=70128164. Version: 2010. – [Online; Stand 16. Mai 2011]

A. Grafiken und Tabellen

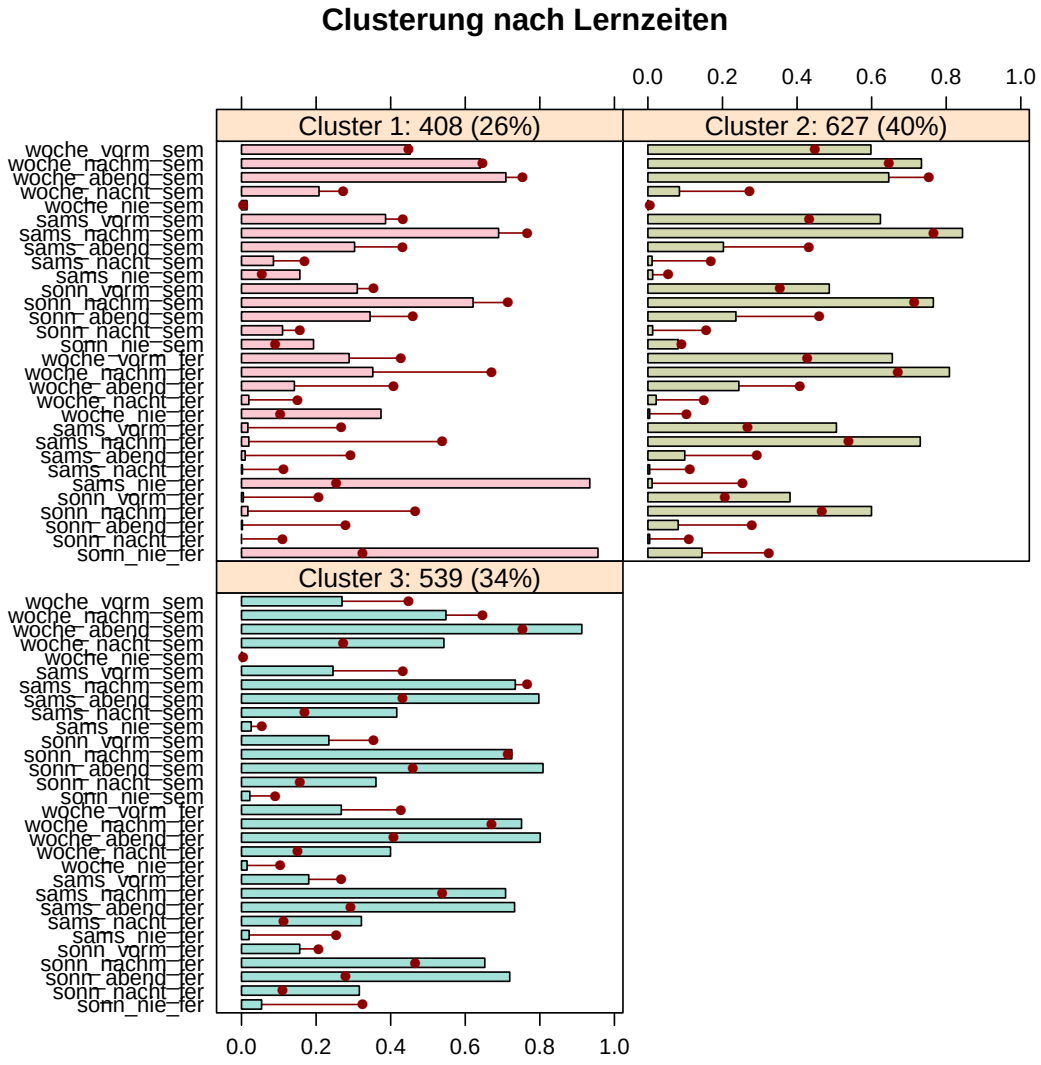

Abbildung A.1.: Darstellung der Zentroide für die Clusterung nach Lernzeiten

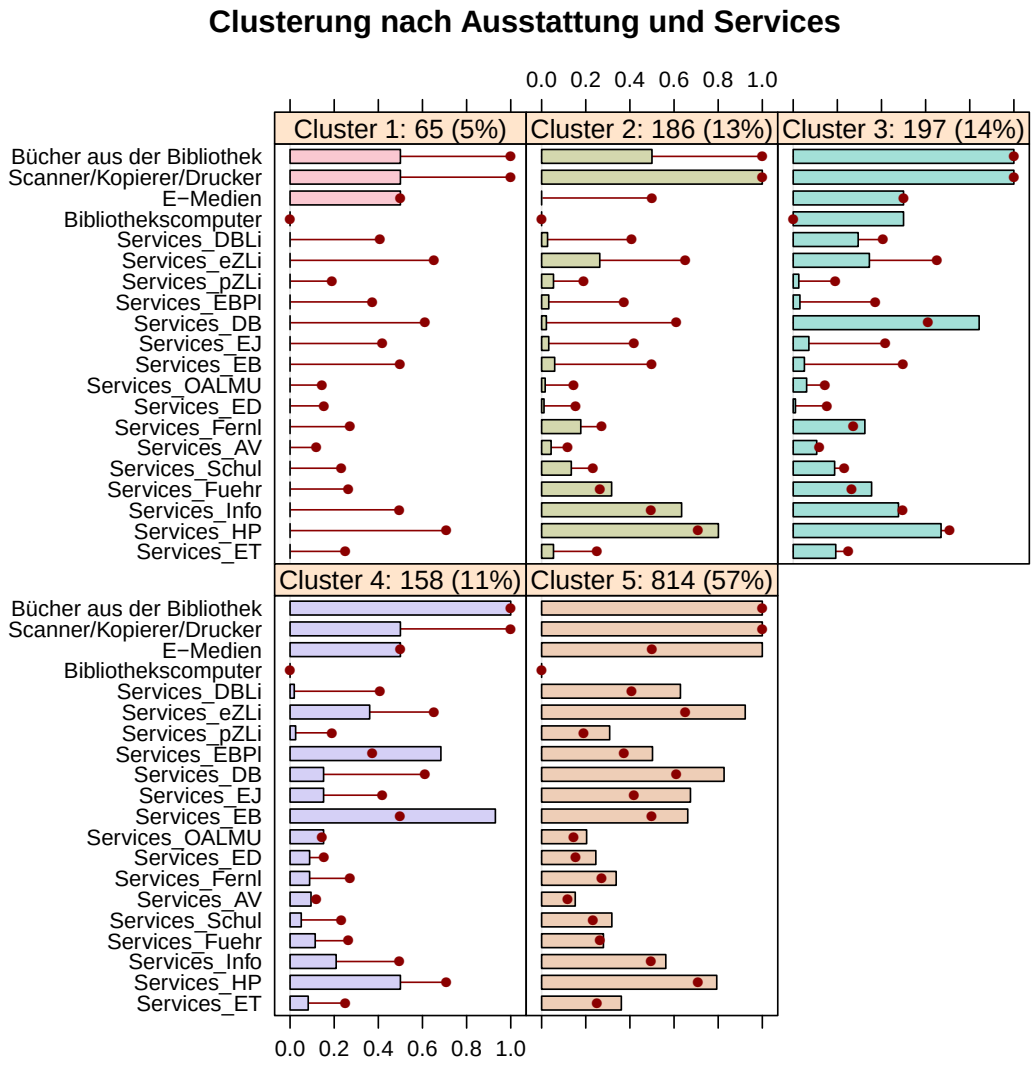

Abbildung A.2.: Darstellung der Zentroide für die Clusterung nach Ausstattung und Services

Abkürzung	Formulierung im Fragebogen
OPAC	Online-Katalog/OPAC
DBLi	Liste der verfügbaren Datenbanken (DBIS)
eZLi	Liste der verfügbaren elektronischen Zeitschriften (EZB)
pzLi	Liste der verfügbaren Print-Zeitschriften (ZDB)
EBPi	E-Book-Plattform
DB	Datenbanken
EJ	E-Journals
EB	E-Books
OALMU	Open Acces LMU
ED	E-Dissertationen
Ausl	Ausleihe
Fernl	Fernleihe
AV	Anschaffungsvorschlag
NeuLi	Neuerwerbungsliste
PrNu	Präsenznutzung (Nutzung nicht ausleihbarer Bücher in der Bibliothek
Schul	Schulungen
Fuehr	Führungen
Info	Infomaterial (Flyer/Aushänge)
HP	Homepage
ET	E-Tutorials

Tabelle A.1.: Antwortmöglichkeiten für die Frage zu den genutzten Services

Abkürzung	Hauptfachgebiet
GeiWi	Geisteswissenschaften
MaNa	Mathematik und Naturwissenschaften
Med	Medizin und Gesundheitswissenschaften
ReWiSo	Rechts-, Wirtschafts- und Sozialwissenschaften

Tabelle A.2.: Abkürzung der Hauptfachrichtungen

Abkürzung	Formulierung im Fragebogen
zuHa	zu Hause
unte	unterwegs
Fak	Fakultät/Institut (ohne Bibliothek)
FachB	Fachbibliothek
UBZ	Universitätsbibliothek/Zentrale
UBS	Universitätsbibliothek/Studentenbibliothek
StaBi	Bayerische Staatsbibliothek
UBTu	Universitätsbibliothek der Technischen Universität München
MueBi	Münchener Stadtbibliothek

Tabelle A.3.: Antwortmöglichkeiten für die Frage zu den Lernorten

Abkürzung	Formulierung im Fragebogen (Zu welchen Zeiten lernen Sie während des Semesters?)
woche_vorm_sem	[Wochentags] [Vormittags(08-12 Uhr)]
woche_nachm_sem	[Wochentags] [Nachmittags(12-18 Uhr)]
woche_abend_sem	[Wochentags] [Abends(18-22 Uhr)]
woche_nacht_sem	[Wochentags] [Nachts(22-08 Uhr)]
woche_nie_sem	[Wochentags] [Nie]
sams_vorm_sem	[Samstags] [Vormittags(08-12 Uhr)]
sams_nachm_sem	[Samstags] [Nachmittags(12-18 Uhr)]
sams_abend_sem	[Samstags] [Abends(18-22 Uhr)]
sams_nacht_sem	[Samstags] [Nachts(22-08 Uhr)]
sams_nie_sem	[Samstags] [Nie]
sonn_vorm_sem	[Sonntags] [Vormittags(08-12 Uhr)]
sonn_nachm_sem	[Sonntags] [Nachmittags(12-18 Uhr)]
sonn_abend_sem	[Sonntags] [Abends(18-22 Uhr)]
sonn_nacht_sem	[Sonntags] [Nachts(22-08 Uhr)]
sonn_nie_sem	[Sonntags] [Nie]

Tabelle A.4.: Abkürzungen der Lernzeiten (Semester)

Abkürzung	Formulierung im Fragebogen (Zu welchen Zeiten lernen Sie während der Semesterferien?)
woche_vorm_fer	[Wochentags] [Vormittags(08-12 Uhr)]
woche_nachm_fer	[Wochentags] [Nachmittags(12-18 Uhr)]
woche_abend_fer	[Wochentags] [Abends(18-22 Uhr)]
woche_nacht_fer	[Wochentags] [Nachts(22-08 Uhr)]
woche_nie_fer	[Wochentags] [Nie]
sams_vorm_fer	[Samstags] [Vormittags(08-12 Uhr)]
sams_nachm_fer	[Samstags] [Nachmittags(12-18 Uhr)]
sams_abend_fer	[Samstags] [Abends(18-22 Uhr)]
sams_nacht_fer	[Samstags] [Nachts(22-08 Uhr)]
sams_nie_fer	[Samstags] [Nie]
sonn_vorm_fer	[Sonntags] [Vormittags(08-12 Uhr)]
sonn_nachm_fer	[Sonntags] [Nachmittags(12-18 Uhr)]
sonn_abend_fer	[Sonntags] [Abends(18-22 Uhr)]
sonn_nacht_fer	[Sonntags] [Nachts(22-08 Uhr)]
sonn_nie_fer	[Sonntags] [Nie]

Tabelle A.5.: Abkürzungen der Lernzeiten (Semesterferien)

B. R Code

B.1. Computational Details

Die folgende Aufzählung beschreibt die verwendete Version von R, sowie sämtliche für Berechnungen und Darstellungen benötigten Pakete mit ihrer Versionsnummer.

- R Version 2.12.0 (2010-10-15), `x86_64-pc-mingw32` (R Development Core Team 2010)
- Locale: `LC_COLLATE=German_Germany.1252`, `LC_CTYPE=German_Germany.1252`, `LC_MONETARY=German_Germany.1252`, `LC_NUMERIC=C`, `LC_TIME=German_Germany.1252`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, stats4, utils
- Other packages: flexclust 1.3-1 (Leisch 2006), lattice 0.19-13 (Sarkar 2008), modeltools 0.2-17 (Hothorn u. a. 2010), xtable 1.5-6 (Dahl 2009)
- Loaded via a namespace (and not attached): tools 2.12.0

B.2. Funktionen

An dieser Stelle wird der Code der Funktion `erzeugeFamily`, mit den darin enthaltenen Funktionen `iDistMixed` und `iCentMixed` beschrieben. Die Funktion `erzeugeFamily` generiert mithilfe der Funktion `kccaFamily` aus dem Paket **flexclust** ein Objekt vom Typ „kccaFamily“, das für die Durchführung einer Clusteranalyse mit gemischten Skalentypen benötigt wird. Dazu müssen der Funktion durch die Parameter `o` und `b` die Spalten, in denen sich ordinale, beziehungweise binäre Variablen befinden, als Vektoren übergeben werden. Die Vergabe von Gewichten ist optional.

```
erzeugeFamily<-function(o, b, ordgewichte=1, bingewichte=1){
  ordinalespalten<-o
```

```
  binärespalten<-b


  iDistMixed<-function(x, centers){
    #aufteilen des Datensatzes und der Zentroide
    xord<-x[,ordinalespalten, drop=F]
    xbin<-x[,binärespalten, drop=F]
    centersord<-centers[,ordinalespalten, drop=F]
    centersbin<-centers[,binärespalten, drop=F]
    #Berechnen der Distanzen jeweils getrennt
    retOrd<- distManhattan(xord,centersord)/length(ordinalespalten)
    retJac<- distJaccard(xbin,centersbin)
    ret<-(ordgewichte*retOrd+bingewichte*retJac)/(ordgewichte+bingewichte)
    rownames(ret)<-rownames(x)
    colnames(ret)<-rownames(centers)
    ret
  }

  iCentMixed<-function (x){
    xord<-x[,ordinalespalten, drop=F]
    xbin<-x[,binärespalten, drop=F]
    centord<-centMedian(xord)
    centbin<-centMean(xbin)
    cent<-c(centord,centbin)
    cent
  }

  retfamily<-kccaFamily(dist=iDistMixed, cent=iCentMixed,
    preproc= function(x) as.matrix(x))
  retfamily
}
```

Die hier definierte Funktion `iDistMixed` beschreibt die Berechnung einer Distanzmatrix zwischen allen Beobachtungen und Zentroiden mithilfe der Funktionen `distManhattan` und `distJaccard` aus dem Paket **flexclust**. Die hier definierte Funktion `iCentMixed` beschreibt die Berechnung des Zentroids einer Gruppe von Beobachtungen mithilfe der Funktionen `centMedian` und `centMean` aus dem Paket **flexclust**.

B.3. Beispielhafter Aufruf einer Clusteranalyse

Der Code für die Reproduktion sämtlicher Analysen befindet sich auf der CD, die dieser Arbeit beiliegt. Die für die weitere Bearbeitung notwendigen Datenobjekte werden durch Ausführen der Skript-Datei „Datenmanipulation.r“ erzeugt. Nun kann mit folgendem Code die Clusteranalyse für Abschnitt 4.4 durchgeführt werden.

```
#Zuerst Datenobjekt generieren
#Dazu Ausstattung ausschneiden
Ausstattung.teil<-Ausstattung.trans[,c("Infrastruktur_BibBü",
 "Infrastruktur_Scan","Infrastruktur_E_hyph_Med",
 "Infrastruktur_PCB")]
#schönere Namen vergeben
names(Ausstattung.teil)<-c("Bücher aus der Bibliothek",
"Scanner/Kopierer/Drucker", "E-Medien",
  "Bibliothekscomputer")

#Services ausschneiden
Services.teil<- Services[,-c(1,11,14,15)]

#Daten zusammenfügen
AusstServ<-cbind(Ausstattung.teil,Services.teil)
#1574 Beobachtungen

#NA bereinigen
AusstServ<-na.omit(AusstServ)
# 1420 Beobachtungen

#kccaFamily objekt für diese Analyse erstellen
AusstServ.fam<-erzeugeFamily(o=1:4,b=5:20,
 ordgewichte=1, bingewichte= 4)

#kcca
set.seed(260788)
AusstServ.analysen<-stepFlexclust(AusstServ, k=2:20,
 nrep=100, family=AusstServ.fam)
```

Mit dem Objekt „AusstServ.analysen“ liegt dann ein Objekt vor, in dem die Ergebnisse der Clusteranalyse für die verschiedenen Clusteranzahlen gespeichert sind. Dieses wird im folgenden Code verwendet, um den Screeplot, sowie Shadow-Plot und Nachbarschaftsgraph für die Lösung mit fünf Clustern zu erstellen.

```
#Screeplot erstellen
plot(AusstServ.analysen, type="l",
 main="Scree-Plot für Clusterung nach \n Ausstattung und Services",
 xlab="Anzahl der Cluster",
 ylab="Summe der Distanzen innerhalb der Cluster")

#untersuchtes objekt extra abspeichern
AS<-AusstServ.analysen[[4]]

#Zentroide und Clustergrößen darstellen
barchart(AS, origin=0, main="Clusterung nach Ausstattung und Services")

#PCA für Projektion durchführen
names(AusstServ)<-names(AS@xcent)
AusstServ.pca<-prcomp(AusstServ)


#neighbourhood und shadowplot für 5 cluster lösung
plot(AS, data=AusstServ, project=AusstServ.pca, points=FALSE,
xlab="1. Hauptkomponente", ylab="2. Hauptkomponente", main="")

plot(shadow(AS))
```

C. Inhalt der CD

Dieser Arbeit liegt eine CD mit folgendem Inhalt bei:

- die Arbeit in elektronischer Form („BA Eifler.pdf“)
- der Praktikumsbericht („Praktikumsbericht.pdf“)
- eine Anleitung zur Verwendung der Daten und zur Reproduktion der Analysen („Anleitung.pdf“).
- Der Ordner „Daten und Analysen“ mit folgender Inhaltsstruktur:
  - Analysen:
    In diesem Ordner befinden sich weiter Ordner, die jeweils den R-Code für die Analysen, sämtliche Grafiken im PDF-Format und ein RData-Objekt mit dem kcca-Objekt der jeweiligen Clusteranalyse beinhalten.
    - Ausstattunswünsche und Services
    - Lernorte
    - Lernzeiten
    - Services
  - Daten und Funktionen:
    In diesem Ordner befinden sich die Ausgangsdaten in Form eines RData-Objekts. R-Code für die Manipulation der Daten und benötigte Funktionen. Ausserdem befinden sich in diesem Ordner zu jeder Code-Datei eine Beschreibung im ODF-Format.
  - Datenbeschreibung:
    In diesem Ordner befinden sich R-Code und Grafiken im PDF-Format zum Kapitel 3.

Danksagung

Frau Prof. Dr. Bettina Grün danke ich für die Überlassung des Themas und für ihre Beratung und Unterstützung bei der Erstellung dieser Arbeit.

Dr. Antje Michel und Medea Seyder von der UB, Robert Pietsch und Malte Schierholz danke ich für die gute Zusammenarbeit während des Statistischen Praktikums, in dem der Datensatz entstanden ist.

Bei den Mitarbeitern der Universitätsbibliothek München bedanke ich mich für die gute Zusammenarbeit und das Überlassen der Daten.

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 31. Juli 2011

.....

(Fabian Eifler)

CD (siehe Anhang C)